Small RNA studies in *Drosophila melanogaster*, *Stylophora pistillata* and *Symbiodinium sp.*

Yi Jin LIEW

A dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy

Trinity Hall

Cambridge

September 2012

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. No part of this thesis has previously been or currently is being submitted for any degree, diploma or other qualification at any other University.

This thesis complies with the regulations set down by the Degree Committee of the School of Biological Sciences.

Signed: .....

Date: 20/09/2012

Acknowledgements

This thesis would not have come to fruition without the guidance of my supervisor Gos Micklem. Thank you for patiently discussing the minutiae of experiments and algorithms with me, as well as encouraging me when my experiments failed to produce anything meaningful. I also owe a huge debt of gratitude to Adrian Carr for the endless stream of advice that helped shape my computational results into what it is today.

This work would have been more sparse if not for the two fruitful collaborations I have had. To Venkat and Julian from University of Glasgow, thanks for providing the smaller tissues for my miRNA assays. To Chris, Mani, Till and other members from the Voolstra Lab in KAUST, thanks for having me over in Saudi Arabia for six weeks, and for giving me my first snorkeling experience. Mani, thanks also for helping me out in the lab, and essentially giving me free reign in there.

The time I spent in Cambridge for my undergraduate and graduate years would not have been that educational or enjoyable without the company of my close friends. To my labmates Antje, Weng Khong and Lalitha, it was great being able to bounce ideas off you freely. Jason, Chris, Jia Hui, Hong King, Siew Kit, Jin Yang, Jit Ern and many others, thanks for all the intellectual discourses and general banter we have had over dinners, and for the (board) gaming sessions that invariably followed after our appetites were sated.

I would also like to thank the Jardine Foundation for funding my undergraduate studies, and the Cambridge Commonwealth Trust and Trinity Hall for my graduate studies. I am forever grateful for the opportunities opened up to me as a result of their benevolence.

Lastly, this thesis is dedicated to my parents and brother: your endless stream of support and encouragement has made this possible. Words cannot describe how much this means to me.

Abstract

Small non-coding RNAs such as microRNAs (miRNAs), small interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs) play a big role in regulating gene expression in cells. In my work, I focus primarily on miRNAs, which represses the expression of the mRNA targets post-transcriptionally.

For Drosophila melanogaster, I predicted the tissue-specific expression of several miRNAs based on the expression levels of the predicted mRNA targets in those tissues. The computational predictions are then followed up by quantitative PCR validation of miRNA expression levels in dissected fly tissues.

For *Stylophora pistillata* (a species of coral found in the Red Sea) and *Symbiodinium sp.* (a photosynthetic, symbiotic algae present in the coral cell), my collaborators and I strived to study the genome, transcriptome and proteome of both organisms. At present, there is another coral genome available — from *Acropora digitifera* — but the large evolutionary distance between both corals (about 240 million years apart) warrants in-depth study of our coral of interest. On the other hand, our Symbiodinium genome will be the first of its kind for any dinoflagellate.

My role in the project was to investigate the small RNAome of both organisms via small RNAseq. As the presence of a thick cell wall in *Symbiodinium sp.* poses a unique challenge to RNA extraction, and compounded by the dearth of literature regarding RNA extraction from the dinoflagellate, we optimised a procedure that consistently produced high quality RNA for downstream sequencing.

From our draft proteome, I showed that the RNA interference (RNAi) machinery is very likely to be present in both organisms. Based on our short RNAseq data, I predicted miRNAs in both organisms. Two of the predicted miRNAs in *S. pistillata* have been identified in other organisms, while all of the predicted miRNAs in *Symbiodinium sp.* were novel.

Contents

List of Figures

List of Tables

List of abbreviations

A. digitifera	*Acropora digitifera*
A. thaliana	*Arabidopsis thaliana*
AGO	Argonaute (protein)
BLAST	Basic Local Alignment Search Tool (software)
BLASTp	protein-protein BLAST
bp	base pair
C. elegans	*Caenorhabditis elegans*
ChIP	Chromatin Immunoprecipitation
D. melanogaster	*Drosophila melanogaster*
dbSNP	Single Nucleotide Polymorphism database (hosted by NCBI)
DNA	Deoxyribonucleic acid
dsRBD	double-stranded RNA Binding Domain
dsRNA	double-stranded RNA
ENCODE	ENCyclopedia Of DNA Elements
EST	Expressed Sequence Tag
H. sapiens	*Homo sapiens*
iTOL	interactive Tree Of Life (software)
KAUST	King Abdullah University of Science and Technology (Saudi Arabia)
KEGG	Kyoto Encyclopaedia of Genes and Genomes
miRISC	miRNA-Induced Silencing Complex
miRNA	MicroRNA
mRNA	messenger RNA
MTase	methyltransferase
N. vectensis	*Nematostella vectensis*
ncRNA	non-coding RNA
nt	nucleotide

P. tetraurelia	*Paramecium tetraurelia*
PAZ	Piwi-Argonaute-Zwille (protein domain)
PCR	Polymerase Chain Reaction
PhyML	Phylogenetic estimation using Maximum Likelihood (software)
piRNA	Piwi-interacting RNA
pre-miRNA	precursor-miRNA
pri-miRNA	primary-miRNA
RAPD	Rapid Amplification of Polymorphic DNA
RefSeq	Reference Sequence database (hosted by NCBI)
RIN	RNA Integrity Number
RISC	RNA-Induced Silencing Complex
RNA	Ribonucleic acid
RNAa	RNA activation
RNase	Ribonuclease
RNA-seq	RNA sequencing
rRNA	ribosomal RNA
RT-qPCR	Real Time-quantitative PCR
S. pistillata	*Stylophora pistillata*
S. pombe	*Schizosaccharomyces pombe*
saRNA	small activating RNA
siRISC	small interfering RNA-Induced Silencing Complex
siRNA	Small interfering RNA
SNP	Single Nucleotide Polymorphism
snRNA	small nuclear RNA
sRNA	small RNA
Symbiodinium sp.	Species of genus *Symbiodinium*
T. thermophila	*Tetrahymena thermophila*
tRNA	transfer RNA
UTR	Untranslated Region

Chapter 1

Introduction

1.1 The importance of non-coding RNA in metazoans

The precise relationship between genome size, number of genes and biological complexity still inspires debates in biological circles today. The notion of biological complexity is notoriously hard to define, but broadly, it refers to the metabolic and developmental complexity of the organism (Taft *et al.*, 2007).

With the advent of powerful and improved sequencing techniques, many organisms have had their genomes sequenced: as of the time of writing, over 5,500 genomes have been sequenced (`http://www.ebi.ac.uk/genomes/wgs.html`). Prior to the burst of sequence data, it was believed that organismal complexity depended on the number of genes (Bird, 1995), but not total genome size due to the presence of large swaths of inert non-coding regions.

However, recent experiments have called into question the view that non-coding sequences are inert in genomes. A large portion of the non-coding human genome had been found to be transcribed — not just introns present within genes, but as antisense transcripts to functional genes and intergenic transcripts as well (Frith *et al.*, 2005). Experiments on other model organisms have revealed that at least 85% of the *Drosophila* genome and 70% of the mouse genome is transcribed; in yeast, a vast majority of its genome is transcribed as well (Mattick, 2007). There exists a correlation between the proportion of non-coding RNA

(ncRNA) in an organism's genome to its complexity, consistent across classes of organisms, such as prokaryotes and eukaryotes, single cell and simple multicellular organisms, and invertebrates and vertebrates (Taft *et al.*, 2007).

As it can be shown that a major output of metazoan genomes is ncRNAs, it seems likely that many of these RNA serve regulatory functions. Well-characterised longer ncRNAs include XIST, which is involved in silencing the extra X chromosome in females (Brown *et al.*, 1991) and HOTAIR, which directs the specialisation of skin cells (Rinn *et al.*, 2007); smaller ncRNAs include large families of RNAs, such as siRNAs (short interfering RNAs), miRNAs (microRNAs) and piRNAs (Piwi-interacting RNAs). An advantage held by RNAs over proteins in directing gene regulation is their sequence specificity — RNAs can direct precise interactions over shorter stretches of nucleotides more efficiently than can be achieved by proteins (Mattick, 2007).

As my work focuses primarily on short noncoding RNAs, especially miRNAs, this chapter will detail miRNAs in depth, followed by brief descriptions of siRNAs, piRNAs, and a class of RNAs that activate (instead of repress) gene expression.

1.2 microRNAs: a brief history

microRNAs (abbreviated as miRNAs) is a class of small non-coding RNAs of ∼22 nt in length which regulate gene expression (Lee and Ambros, 2001). miRNA made its debut 17 years ago, when Lee and colleagues identified a non-coding gene, lin-4, that was involved in the negative regulation of LIN-14 protein levels in *C. elegans* (Lee *et al.*, 1993). The term "microRNA" was coined in 2001 in a series of papers that were published in Science Lagos-Quintana *et al.* (2001); Lau *et al.* (2001); Lee and Ambros (2001). As the first few miRNAs were temporally expressed, these small RNAs were also known as "stRNAs" (small temporal RNAs) (Lagos-Quintana *et al.*, 2001), but usage of this abbreviation has been largely superseded by the more general "miRNA".

Interestingly, one of the first few identified miRNAs was discovered to be fairly well conserved across metazoan model organisms — in fact, this discovery predates the official

naming of this class of small RNAs. Although lin-4 appeared to be specific to *C. elegans*, the second miRNA discovered, *C. elegans* let-7, was identified in a wide range of metazoans, such as vertebrates, ascidians, hemichordates, molluscs, annelids and arthropods. Also, the expression of these let-7 homologues is similar to that in *C. elegans* (Pasquinelli *et al.*, 2000). A year later, several other examples of common miRNAs that are present in worms, flies, humans, mice, fish and frog were published (Lagos-Quintana *et al.*, 2001).

The increasing use of next generation sequencing techniques has greatly increased the rate at which new miRNA genes are identified. As RNAseq is more sensitive and specific than hybridisation-based methods at quantifying RNA expression (Wang *et al.*, 2009), the former method is able to pick up miRNAs that are expressed at low levels. Currently, miRNAs have been identified in 193 species that span across major kingdoms of life: animals, plants, amoeba and viruses (from miRBase v19, August 2012) (Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006, 2008; Kozomara and Griffiths-Jones, 2011).

1.3 miRNA biogenesis

Currently, the molecular mechanism underlying the generation of miRNAs have been fairly well studied. The biogenesis of most miRNAs can be roughly separated into four processes: the initial transcription of miRNAs as long precursors called pri-miRNAs (primary-miRNAs), the processing of the longer pri-miRNAs into the shorter hairpin precursors called pre-miRNAs (precursor-miRNAs) in the nuclei, the export of the pre-miRNAs into the cytoplasm, and lastly, the processing of the double-stranded pre-miRNAs into single-stranded mature miRNAs in the cytoplasm. An overview of these processes, which are detailed as separate subsections below, can be found in Figure 1.1.

Most pre-miRNAs give rise to two miRNA sequences, one from each arm of the hairpin. As per miRNA naming conventions, the predominant sequence is assigned a numerical identifier with a "miR-" prefix (e.g. miR-100), while the less predominant one has an added asterisk at the end (e.g. miR-100*) (Ambros *et al.*, 2003; Griffiths-Jones, 2004).

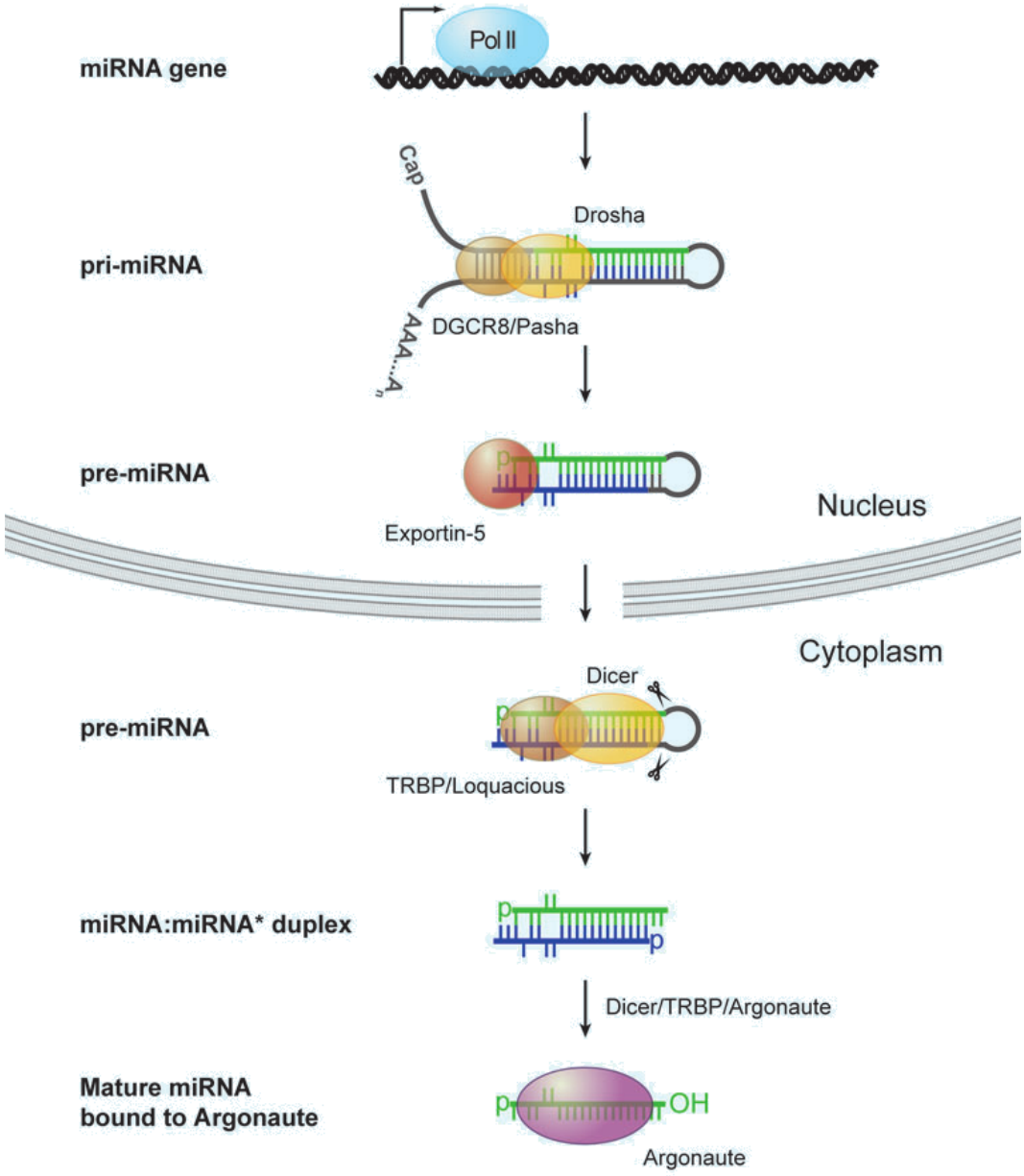


Figure 1.1: A general overview of miRNA biogenesis in metazoans. The crucial proteins involved in the transcription and maturation of miRNAs are also shown (Bushati and Cohen, 2007).

1.3.1 Initial transcription of pri-miRNAs

The transcription of pri-miRNAs is carried out by RNA polymerase II, followed by capping and polyadenylation of the transcripts (Kim, 2005). Initially, it was believed that miRNAs were transcribed by RNA polymerase III, similar to the transcription of the small and ubiquitous tRNAs and U6 snRNA. However, some circumstantial evidence from early experiments casted doubt on the initial belief. It was found that pri-miRNAs can be of several kilobases long, and could contain stretches of four uracils, which would have terminated transcription by RNA polymerase III (Lee *et al.*, 2004a).

On top of that, early EST analyses found chimeric transcripts that contained miRNA sequences and fragments of adjacent mRNAs on the same transcript, suggesting that RNA polymerase II is responsible for the transcription of these transcripts (Smalheiser, 2003), and this was later confirmed by work done by other groups. While some animal miRNAs are located on individual transcription units, many other miRNAs are produced from transcription units that contain more than one product (Bartel, 2004). The transcript might contain clusters of distinct miRNAs, or it may encode a miRNA and protein. Examples of the former include the miR-309 cluster in flies (Biemar *et al.*, 2005) and the miR-17-92 cluster in humans (Hayashita *et al.*, 2005); the latter is exemplified by miR-7, which resides in the *Drosophila* gene *bancal* (Li and Carthew, 2005). It is thought that many animal miRNAs arise from the accumulation of nucleotide sequence changes, and not from gene duplication — if the new miRNA sequence appears in an existing transcriptional unit, the miRNA will be expressed, despite not having enhancer and promoter sequences (Carthew and Sontheimer, 2009).

Lastly, in *Drosophila*, the insertion of RNA polymerase II transcriptional enhancers upstream of *bantam* increased the level of mature *bantam* miRNA (Brennecke *et al.*, 2003); in *C. elegans*, the temporal regulation of let-7 expression was regulated by an enhancer element known as TRE (temporal regulatory element) (Johnson *et al.*, 2003). These evidence, while indirect, strongly suggests that RNA polymerase II is involved in the biogenesis of miRNAs.

Direct evidence proving the involvement of RNA polymerase II in the transcription of

pri-miRNAs was achieved using GST-tagged (glutathione-S-transferase-tagged) eIF4E that selectively binds to RNA with 7-methyl guanosine cap structures in HeLa cells. Lee *et al.* (2004a) found that these caps were found in seven randomly chosen pri-miRNAs sequences, but not in their corresponding pre- or mature miRNAs. To investigate the presence of poly(A) tails on the pri-miRNAs, beads with oligo-dT were used to extract polyadenylated RNA, and the resulting bound and unbound fractions subjected to an RNase protection assay or RT-PCR. Both approaches showed that the probed pri-miRNAs could be found in both the bound and unbound fraction, indicating that pri-miRNAs do possess poly(A) tails. Also, the production of miRNAs is halted to a large degree by subjecting cell extracts to alpha-amanitin at levels that block RNA polymerase II activity. Despite blocking RNA polymerase II, electrophoretic gels revealed the presence of trace pri-miRNAs in the cell extracts. It is possible that a minor fraction of pri-miRNAs are actually transcribed by other polymerases, or it might be that the processing of pri- to pre-miRNAs is slower than transcription of pri-miRNAs, explaining the faint presence of the probed miRNA several hours after alpha-amanitin was added.

This canonical view of pri-miRNA transcription did not hold for long. A few years later, there was evidence that linked RNA polymerase III to the transcription of human miRNAs in the chromosome 19 miRNA cluster (Borchert *et al.*, 2006), and more recently, the viral murid herpesvirus 4 miRNA as well (Diebel *et al.*, 2010). In both studies, the biogenesis of the miRNAs was immune to the presence of alpha-amanitin, which was added to inactivate RNA polymerase II. Although Borchert *et al.* (2006) estimated that the transcription of over 20% of human miRNAs is under the control of RNA polymerase III, it remains to be seen whether this estimation is valid in the other model organisms as well.

1.3.2 Conversion of pri-miRNAs to pre-miRNAs

After transcription, these pri-miRNAs need to undergo further nuclear processing into ∼70 nt stem-loop precursors, known as pre-miRNA. This processing is dependent on the pri-miRNA folding into a stem-loop structure, which consists of a ∼33 nt imperfectly-paired stem, with

a terminal loop and flanking segments (Bartel, 2004). Excision of this imperfectly-paired stem occurs in the nucleus, catalysed by Drosha in animals (Lee *et al.*, 2003) and DCL1 (Dicer-like 1) in plants (Kim, 2005).

The activity of Drosha was first elucidated by Lee *et al.* (2003). Using pri-miR-30a from HEK293T (human embryonic kidney) cells, pre-miR-30a was produced in an *in vitro* system. The pre-miR-30a was then subjected to RT-PCR, cloning and sequencing, which resulted in the identification of pre-miR-30a as a stem-loop of 63 nucleotides. The structure of pre-miR-30a had two interesting features. Firstly, it had a 2 nt overhang at the 3' end, which is characteristic of products of RNase III-mediated cleavage. This was further confirmed by experiments published in the same paper. Mutations that disrupted the stem-loop structure drastically reduced the efficiency of processing the pri-miRNAs, which is expected as RNase III enzymes bind specifically to dsRNA (double-stranded RNA); also, similar to other members of the RNase III family, the catalytic reaction of Drosha in vitro required the presence of divalent cations ($MgCl_2$ was added to the buffer in the experiment). Secondly, the termini of pre-miR-30a was identical to both the mature miR-30a and miR-30a*, which indicated that the cleavage reaction by Drosha predetermines one end of the mature miRNAs (Lee *et al.*, 2003).

However, for proper functioning of Drosha, it has been shown that Pasha, a double-stranded RNA binding protein, stably associates with the ribonuclease in a complex termed the Microprocessor complex (Denli *et al.*, 2004). The role of Pasha can be likened to that of a molecular ruler — Pasha binds to the junction between the dsRNA pri-miRNA stem and the flanking ssRNA sequences, allowing Drosha to rely on the position of Pasha to cut at a location one helix turn (11 bp) away from the junction between the dsRNA pri-miRNA stem and the flanking ssRNA sequences (Han *et al.*, 2006). Structurally, Drosha is large protein ($\sim$160 kDa in humans) that contains two RNase III domains and a double-stranded RNA-binding domain (dsRBD), while Pasha contains two dsRBDs as well. The Microprocessor complex is $\sim$500 kDa in *D. melanogaster* and $\sim$650 kDa in humans (Kim, 2005).

However, there are exceptions to this rule. The splicing of some pri-miRNAs produces

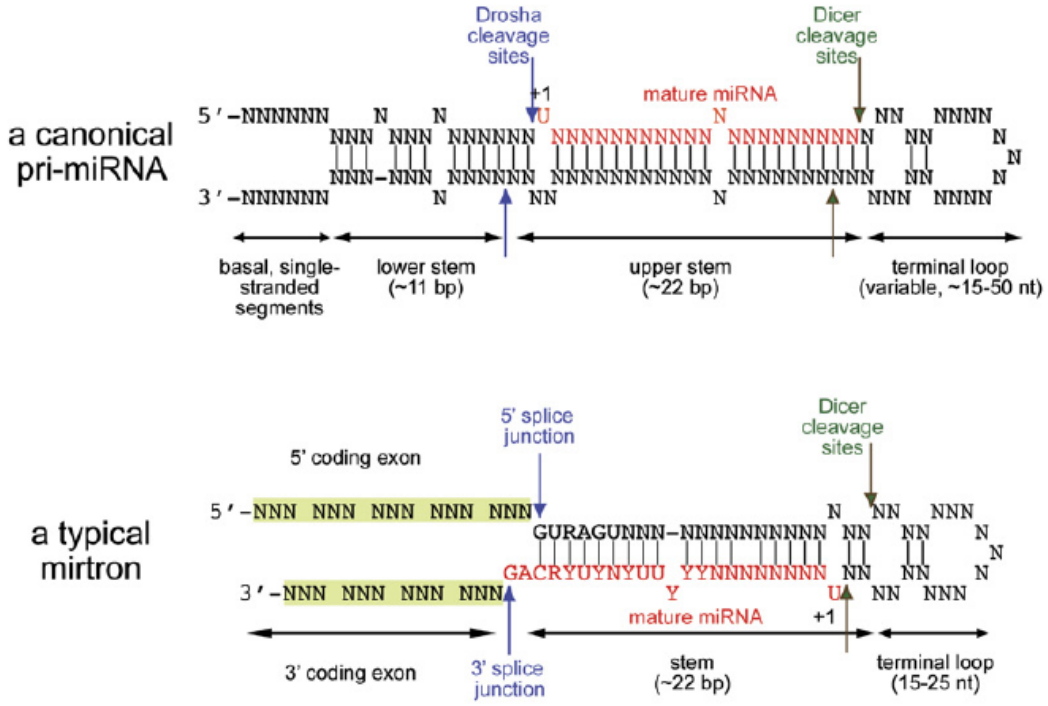


Figure 1.2: A comparison of the analogous processing of conventional pri-miRNAs and mirtrons within coding regions. Both methods produce pre-miRNAs that are recognised by Dicer in the cytoplasm. Illustration from Okamura *et al.* (2007).

introns that mimic the structure of pre-miRNAs. These miRNAs, termed mirtrons, while not common, are found scattered in the animal kingdom. The existence of mirtrons was first predicted from the 454 pyrosequencing of *Drosophila* small RNAs by the Bartel lab — they found 14 short introns that had hairpin structures resembling those of pre-miRNAs. These short introns lack the lower ∼11 bp stem of pri-miRNAs, which indicates that these introns are not processed by the Microprocessor complex. Instead, the hairpin ends of the mirtrons correspond directly to the splice sites (the 5' donor site and the 3' acceptor site), while the AG consensus at the acceptor site mimics the 2 nt 3' overhang of conventional pre-miRNAs after cleavage by Drosha (Okamura *et al.*, 2007). A diagrammatic comparison of mirtrons with conventional pre-miRNAs can be seen in Figure 1.2.

1.3.3 Nuclear export of pre-miRNAs

As Drosha and Dicer are compartmentalised to the nucleus and cytoplasm of the cell respectively, nuclear export of pre-miRNAs is essential to complete the maturation of miRNAs. The export of pre-miRNAs from the nucleus was found to be sensitive to the depletion of nuclear RanGTP, achieved either via the injection of RanGTPase activating protein (Bohnsack *et al.*, 2004) or the inhibition of Ran guanine nucleotide exchange factor (Lund *et al.*, 2004). Peptide mass fingerprinting and immunoblotting demonstrated that the transport of pre-miRNAs across the nuclear membrane is mediated by exportin-5, a nuclear transport receptor (Bohnsack *et al.*, 2004). This observation is corroborated by other experiments that used synthetic siRNAs to downregulate exportin-5 expression in human and *Xenopus* cells, which resulted in the reduction of mature miRNAs in the cytoplasm of the cells (Yi *et al.*, 2003; Lund *et al.*, 2004). Exportin-5 was initially identified as a minor transport factor for tRNAs, but follow-up studies revealed that exportin-5 has a much higher affinity for miRNAs than tRNAs. As the cell does contain many copies of mature miRNAs at any time, it is highly likely that pre-miRNAs are the main cargoes for exportin-5 (Kim, 2005).

It was found that depleting exportin-5 led to the steep reduction of pre-miRNA and mature miRNA levels in the cytoplasm, but interestingly, the levels of pre-miRNAs in the nucleus do not show a converse increase, as would be expected, from the absence of exportin-5. It is possible that pre-miRNAs are stabilised by the physical interaction with exportin-5 — unbound pre-miRNAs are vulnerable to exonuclease function in the nucleus (Yi *et al.*, 2003).

To identify the structural requirements for efficient binding of exportin-5 cargo to its receptor, mutational studies were carried out on pre-miR-30a in HeLa cells. Two factors were of interest: the length of the dsRNA stem, as well as the presence or absence of short overhangs at the termini of the RNA. Variations from the canonical pre-miRNA structure was surprisingly well tolerated. Although the most efficient export by exportin-5 was achieved when the length of the dsRNA stem of the cargo exceeds 16 bp, at the same time having a short overhang at the 3' end, many constructs with blunt ends achieved near wild-type rates

of export. The only factor clearly detrimental to nuclear export is the presence of overhangs of any length at the 5' end of the cargo (Zeng and Cullen, 2004).

1.3.4 Maturation of pre-miRNAs and silencing of mRNA targets

The final step in miRNA biogenesis is the processing of double-stranded pre-miRNAs into the ∼22 nucleotide single stranded mature miRNAs, and the subsequent loading of these mature miRNAs into miRISCs (miRNA-induced silencing complexes) that regulates the post-transcriptional expression of their targets. Both of these processes are carried out by the RISC loading complex, consisting of Dicer, TRBP (Tar RNA binding protein), PACT (protein activator of PKR) and Ago (Argonaute, which also mediates RISC effects on mRNA targets). While TRBP and PACT are not essential for Dicer-mediated cleavage of pre-miRNAs, depletion of both proteins reduces the efficiency of post-transcriptional gene silencing (Winter *et al.*, 2009). On top of recruiting Argonaute to the RISC loading complex (Chendrimada *et al.*, 2005; Lee *et al.*, 2006), it is thought that both TRBP (Chendrimada *et al.*, 2005) and PACT (Lee *et al.*, 2006) stabilise Dicer, but there are conflicting observations — Haase *et al.* (2005) reports that the depletion of TRBP in HEK293 cell lines did not result in an appreciable destabilisation of Dicer.

1.3.4.1 Action of Dicer on pre-miRNAs

To produce mature miRNAs with highly exact ends, another cut has to be made a fixed distance away from the ends produced by Drosha. In animals, this step is mediated by Dicer in the cytoplasm; in plants, DCL1 carries out this step in the nucleus (Kim, 2005). Similar to Drosha, Dicer is a member of the RNase III superfamily of nucleases, first identified in work done by Bernstein *et al.* (2001). There is more variation in the structure of Dicer across organisms than Drosha — most metazoan Dicers have, from the amino to carboxy terminus, a DExD/H ATPase domain (has helicase activity), a DUF283 domain (most likely for dsRNA binding), a PAZ domain (preferentially binds to the 3' single-stranded overhang of miRNAs, acts as a molecular ruler for the nuclease), two RNase III domains (catalyses

dsRNA cleavage) and a dsRBD domain (dsRNA binding domain)(Lee *et al.*, 2003; Dlakic, 2006; MacRae *et al.*, 2006; Carthew and Sontheimer, 2009). Unlike Drosha, Dicer plays roles in the biogenesis of both siRNAs and miRNAs. Some organisms, such as mammals and nematodes, have one Dicer to handle the biogenesis of both small RNAs, while other organisms have multiple types of Dicers (two in *D. melanogaster*, four in *A. thaliana*). The presence of several distinct Dicers usually indicates functional specialisation between them, as exemplified by *Drosophila*: Dicer-1, acting with a dsRNA-binding protein Loqs (Loquacious), is required for miRNA biogenesis, while Dicer-2 is involved in siRNA processing (Tomari and Zamore, 2005; Förstemann *et al.*, 2007; Carthew and Sontheimer, 2009).

In the case of miRNAs with a high degree of complementarity in the hairpin stem, there is an additional layer of processing before the action of Dicer on the pre-miRNAs. Ago makes a single-stranded nick in the middle of the strand opposite the mature miRNA, leaving the mature miRNA sequence intact. This intermediate form, known as the ac-pre-miRNA (Ago2-cleaved pre-miRNA, see Figure 1.3), is recognised by Dicer as readily as other pre-miRNAs. This Ago2-mediated cleavage is thought to facilitate strand disassociation and activation of the miRISC (Winter *et al.*, 2009).

The molecular mechanism of Dicer's nuclease activity was revealed by studying a conserved Dicer protein in *Giardia intestinalis*, a flagellated protozoan parasite inhabiting the small intestine of its mammalian host. *Giardia* Dicer was able to produce short RNAs of 25–27 nt in length, while having just three conserved features on the protein — a PAZ domain and two RNase III domains. The crystal structure of *Giardia* Dicer resembles that of a hatchet, where the RNase III domains form the blade, and the PAZ domain makes up the base of the handle. The RNase III domains and the PAZ domain are directly connected by a long alpha-helix that runs through the handle of the molecule. The length of the small RNAs produced (∼25 nt) directly corresponds to the length of this alpha-helix (∼6.5 nm), suggesting that PAZ acts as a molecular ruler by preferentially binding to the 2 nt 3' ssRNA overhang of the dsRNA, and then directing the RNase III domains to cleave the dsRNA at a specified distance from the bound end (MacRae *et al.*, 2006; Carthew and Sontheimer,

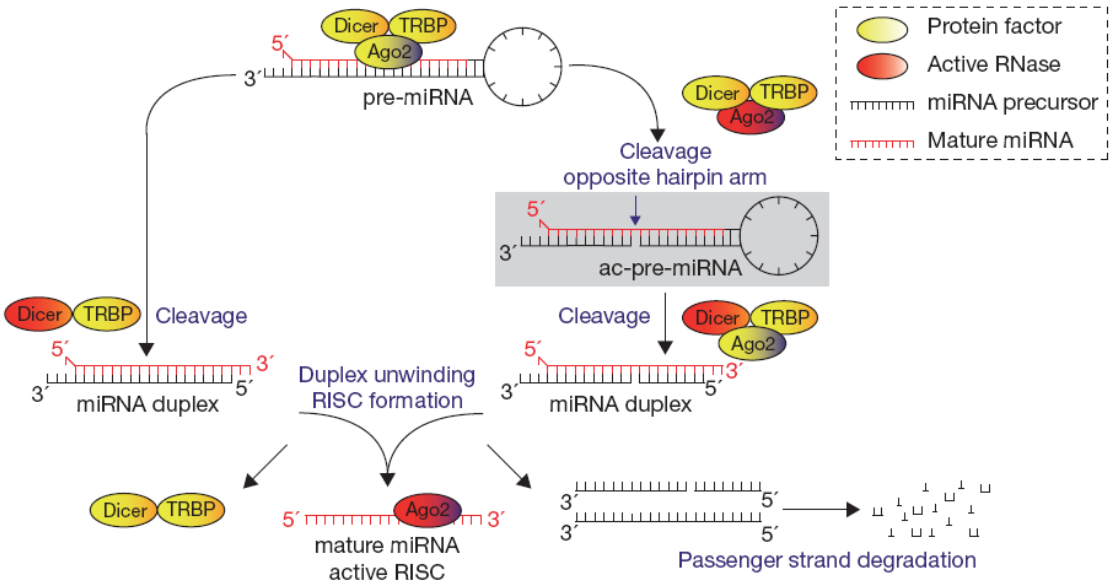


Figure 1.3: Divergence in the pathway of generating mature miRNAs from pre-miRNAs. The intermediate form ac-pre-miRNA is highlighted in grey. Both pathways converge again after mature miRNAs are produced from the respective pathways (Winter *et al.*, 2009).

2009). After cleavage, Dicer and TRBP/PACT dissociate from the RISC loading complex (Winter *et al.*, 2009).

1.3.4.2 miRNA strand selection

As the miRNAs cleaved by Dicer are double-stranded in nature, theoretically speaking, there are two potential forms of mature miRNAs that could arise from this duplex. In reality, similar to cleaved siRNA duplexes, one strand is predominantly incorporated into the miRISC, while the other strand tends to be degraded (Winter *et al.*, 2009).

This asymmetric nature of miRNA loading has been attributed to the different thermodynamic profiles of both ends of the miRNA duplex. In a very comprehensive computational analysis by Khvorova *et al.* (2003), they found that the miRNA strand which is preferentially loaded into miRISC had less stable base pairing at the 5' end, and had low internal stability (in the form of internal mismatches and bulges) about 9–14 bp from the same 5' end. The application of these "loading rules" correctly predicted the accumulation of 20 out of 27 functional mature miRNAs in a biochemical study done by Schwarz *et al.* (2003).

Of the seven remaining miRNAs, which were predicted to accumulate in both the mature miRNA form and the antisense miRNA* form, five have experimental evidence supporting the prediction.

The preferential incorporation of the mature miRNA strand over the antisense miRNA* one bears many similarities to strand selection in siRNAs as well (Schwarz *et al.*, 2003), which is not entirely unexpected as both miRNAs and siRNAs require the use of Dicer enzymes for the generation of their mature form, and the use of Argonaute proteins to support their silencing function (Carthew and Sontheimer, 2009).

1.3.4.3 Post-transcriptional silencing of target mRNAs

After strand selection, the mature miRNAs can either direct translational repression or endonucleolytic degradation of target mRNAs depending on which Ago protein it associates with in the RISC complex (He and Hannon, 2004; Förstemann *et al.*, 2007; Iwasaki *et al.*, 2009). In *D. melanogaster*, there are five Argonaute proteins that form two subclades, named Ago1 and Ago2. Previously, Ago1 and Ago2 were reported to bind to miRNAs and siRNAs respectively (Okamura *et al.*, 2004), but some exceptions to the rule have been discovered — although *Drosophila* miR-277 is produced by Dicer-1, it is loaded into Ago2 (Förstemann *et al.*, 2007). The current prevailing view is that despite having distinct biogenesis pathways for both miRNAs and siRNAs, they participate in a common sorting step after its biogenesis, with their intrinsic structures determining their eventual association with either Ago1 or Ago2 (Förstemann *et al.*, 2007; Iwasaki *et al.*, 2009).

In flies, it has been previously proposed that Ago1 mediates translational repression while miRNA-loaded Ago2 mediates siRNA-like cleavage of its targets (Förstemann *et al.*, 2007). Experiments carried out by Iwasaki *et al.* (2009) showed that Ago2 could also direct translational repression, on top of its better-studied role in inducing target cleavage. Using an *in vitro* assay, the association of miR-277 to Ago1 directed about 8-fold repression of the reporter gene, while Ago2 managed a roughly 2.5-fold repression. It is possible that previous studies, the knocking out of Ago2 was compensated by the translational repression activity

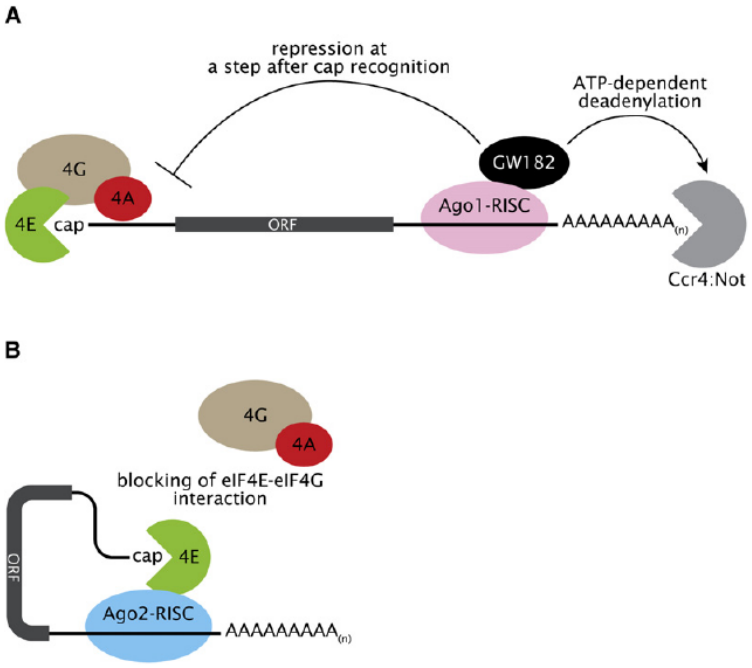


Figure 1.4: Model of Ago1- and Ago2-mediated translational repression. Ago1 has two independent translational repression mechanisms, but both require GW182 to function (Iwasaki *et al.*, 2009).

of Ago1, giving the false impression of Ago2 not being able to repress translation.

However, the molecular mechanisms governing the repression of translation by Ago1 and Ago2 are distinct (see Figure 1.4). Ago1 has two mechanisms to repress translation: by directing ATP-dependent deadenylation of the targeted mRNAs, or by repressing the recognition of the cap structure formed at the 5' end of the mRNAs. Both of these mechanisms require the P-body protein GW182. A C-terminal truncation of the protein allows it to bind to Ago1, but it neither represses the translation of the mRNA targets nor direct Ago1 to P-bodies in *Drosophila* S2 cells. Ago2, however, exerts its repression on translation in a GW182- and ATP-independent manner by blocking the formation of the cap structure. It achieves this by interfering with the interaction between cap proteins eIF4E and eIF4G as mRNA-bound Ago2-RISC complexes have a dramatically enhanced affinity for the eIF4E protein (Iwasaki *et al.*, 2009).

Unlike animal miRNAs, plant miRNAs tend to have near-perfect or perfect complementarity to their targets — it was thus thought that endonucleolytic degradation was its primary, or even exclusive, mode of repression (Jones-Rhoades *et al.*, 2006). Recent experimental work

revealed that the action of miRNAs in plants and animals might not be as different as was initially thought. Despite the highly complementary miRNA-mRNA target binding, analysis done on plant mutants have highlighted translational repression as the "default" mechanism, while perfectly complementary miRNAs may additionally engage in endonucleolytic cleavage of target mRNAs on top of the repression (Brodersen *et al.*, 2008).

1.4 miRNA-mRNA target recognition

Although much is known about the molecular mechanism of miRNA biogenesis in the cell, comparatively less is elucidated about miRNA recognition and binding to its regulatory targets. The first hint about miRNA target recognition was from the experimental work done on *C. elegans* lin-4, the first miRNA identified. Its target, lin-14, had regions in the 3' UTR (untranslated region) that were complementary to the sequence of lin-4. Molecular genetic analyses showed that these sites were essential in repression of lin-14 by lin-4 (Ruvkun *et al.*, 1989; Wightman *et al.*, 1991; Lee *et al.*, 1993). The crucial role of complementarity to the miRNA sequence in the 3' UTR was further reinforced by the discovery of such sites in lin-28 and lin-41, which are targeted by lin-4 and let-7 respectively (Moss *et al.*, 1997; Vella *et al.*, 2004).

Following the discovery of more miRNAs via cloning or computational methods, interest soon shifted to identifying the mRNA targets of these known miRNAs. Target prediction is much more straightforward in plants, as plant miRNAs tend to show near perfect complementarity to their target mRNAs, but not so in metazoan miRNAs (Rhoades *et al.*, 2002). Nevertheless, there are three major principles underlying metazoan miRNA-mRNA targeting that have been used by many computer algorithms to predict mRNA targets for known metazoan miRNAs. These principles appear to be largely consistent across a growing body of experimental data (Bartel, 2009).

Firstly, despite the imperfect binding of metazoan miRNAs to their mRNA targets, the 5' halves of these miRNAs are more conserved than the other half of the miRNA (Lim *et al.*, 2003). Also, for some *Drosophila* miRNAs, the 5' end shows perfect complementarity to its

targets (Lai, 2002). Particular attention was paid to a strongly conserved region centered around nucleotides 2–8 of the miRNA — this ∼7 nt region has been called the miRNA "seed" sequence (Lewis *et al.*, 2003). It has been confirmed by experiments that mutations in the seed region have a significant effect on the regulatory effects of many miRNAs (Doench and Sharp, 2004; Brennecke *et al.*, 2005). However, the experimental evidence should not be taken to mean that the 3' end of miRNAs serve no function at all — Brennecke *et al.* (2005) showed that synthetically-induced mismatches in the 5' region can be compensated by strong 3' pairing. One good example illustrating this non-canonical binding *in vivo* is the zebrafish miR-214 gene. This miRNA is able to repress both *su(fu)* (suppressor of fused) and *disp2* (dispatched homolog 2) via three weak, non-canonical elements that do not have perfect seed matches to miR-214 (Li *et al.*, 2008b).

Secondly, just by virtue of searching the 3' UTR of mRNAs for conserved pairing to known miRNA seed regions, mRNA targets can be predicted fairly accurately (Bartel, 2009). For example, in the TargetScanFly algorithm, the ratio of predicted targets to estimated false positives was calculated to be 3.5:1 in an analysis across five genomes. This ratio was higher if position 1 of the miRNA is constrained to be an adenine, but the restriction also results in a significant drop in sensitivity (Lewis *et al.*, 2005).

Due to the short length of miRNAs, it is not surprising that there are many predicted targets for each miRNA. However, after filtering out targets conserved by chance, the number of targets conserved across organisms remained unexpectedly high. This led to the third conclusion: highly conserved miRNAs have many conserved targets (Brennecke *et al.*, 2005; Lewis *et al.*, 2005). For example, the number of mRNA targets per miRNA was estimated at about 100 in *Drosophila* (Brennecke *et al.*, 2005) and about 400 in mammals (Friedman *et al.*, 2009). This conclusion is supported by experimental work. In one study, miR-1 and miR-124 were introduced into HeLa cells, followed by microarray studies 12 hours later. It was observed that hundreds of mRNAs were downregulated in the cells, shifting the profile of the cells to the tissues that preferentially express that miRNA (miR-1 in muscle and miR-124 in brain) (Lim *et al.*, 2005).

1.5 Other small noncoding RNAs

Although most of my work centers around miRNAs, there are several other classes of non-coding small RNAs that are fairly well studied currently. These classes are explained in the sections below.

1.5.1 Small interfering RNAs

Small interfering RNAs (siRNAs) and miRNAs are the two main classes of small RNAs involved in RNA interference (RNAi). The discovery and understanding of RNAi has revolutionised biology, as evidenced by the awarding of the Nobel Prize for Physiology and Medicine to Andrew Fire and Craig Mello for one of the earliest work done in the field (e.g. Fire *et al.* (1998)).

siRNAs and miRNAs differ mainly in their origins — while miRNAs are wholly endogenous in origin, siRNAs can be both endogenous or exogenous in origin. The endogenous generation of siRNAs differs from that of miRNAs, as shown in Figure 1.5.

The incorporation of siRNAs into siRISCs (siRNA-induced silencing complexes) and the subsequent silencing of mRNA targets are largely similar to the downstream effects of mature miRNAs (which has been previously detailed in Section 1.3.4).

1.5.2 Piwi-interacting RNAs

Piwi-interacting RNAs (piRNAs) are small RNA that were identified through association with Piwi proteins (a subtype of Argonaute proteins) in mammalian testes. In contrast to miRNAs, piRNAs are slightly longer — the typical length being 26–30 nt (Kim, 2006; Brennecke *et al.*, 2007). In invertebrates and vertebrates, there is a bias for uridine to be present as the first nucleotide at the 5' end. The 3' end of piRNAs in several metazoans (e.g. worm (Ruby *et al.*, 2006) and mouse (Kirino and Mourelatos, 2007)) are 2'-O-methylated as well, which is thought to increase the stability of the RNA (Faehnle and Joshua-Tor, 2007).

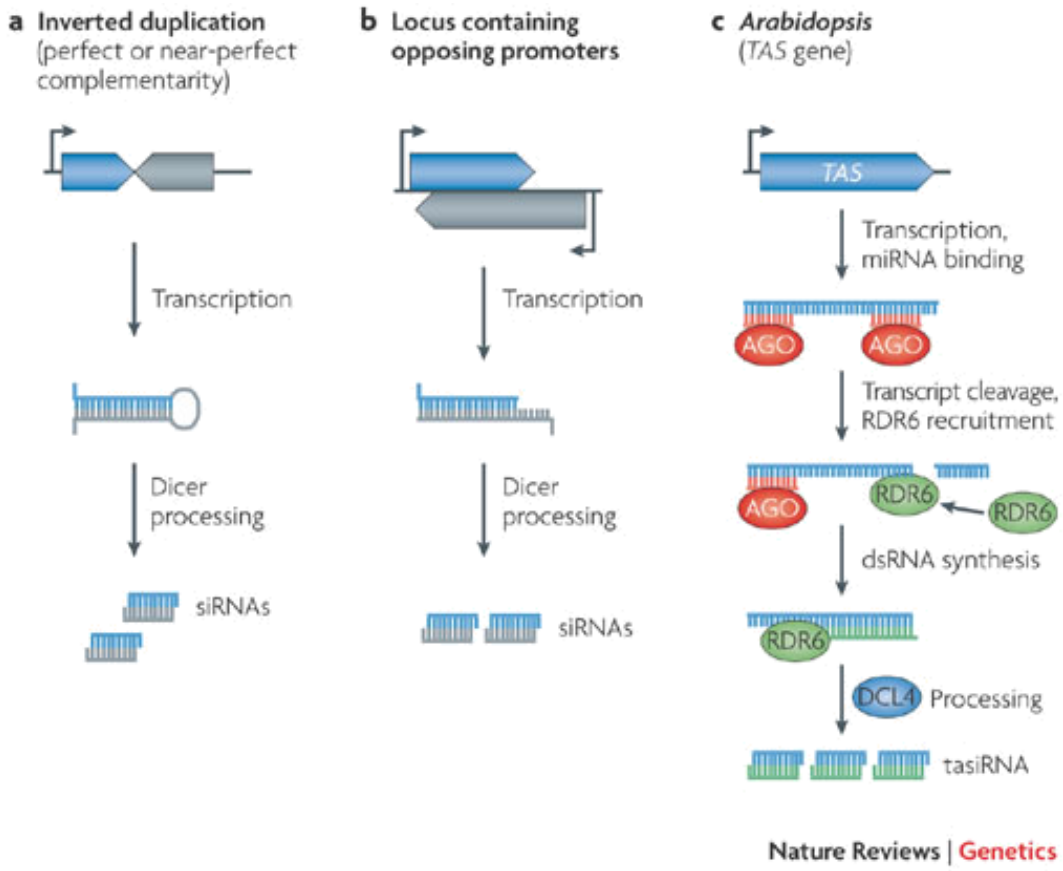


Figure 1.5: Overview of the biogenesis of endogenous siRNAs. In **(a)** and **(b)**, the blue and gray regions are perfect or near-perfect complements of each other. siRNAs are then produced from Dicer processing of the double stranded intermediates. In **(c)**, TAS (trans-acting siRNA) genes in Arabidopsis contain multiple siRNAs lined up one after another. Specific miRNAs induce the cleavage of TAS transcripts at a site upstream of the first siRNA. RDR6 (RNA-dependent RNA polymerase 6) is then recruited to produce the complementary strand, which is then processed by DCL4 (Dicer-like 4) into siRNAs. Illustration adapted from Chapman and Carrington (2007).

Currently, the biogenesis of piRNAs remain poorly understood. In mice, the mapping of piRNA sequences back to the genome reveals a strong strand bias of the sequences, indicating that the biogenesis of piRNAs is likely to involve a long single-stranded precursor (Seto *et al.*, 2007). In addition to the generation of primary piRNAs, Brennecke *et al.* (2007) proposed a "ping-pong" system of piRNA biogenesis: Piwi proteins, after associating with the primary piRNA sequences, is involved in the production of secondary piRNAs that are complementary to the primary piRNA sequences. These secondary piRNAs then guide the same Piwi proteins in directing the generation of more primary piRNA sequences, effectively acting as a positive feedback loop that increases the amounts of both primary and secondary piRNA sequences. However, this "ping-pong" mechanism seems to be restricted to certain organisms — it has been shown that piRNA generation in adult mouse testes is independent of the proteins required in the "ping-pong" model (Beyret *et al.*, 2012).

In mammals and *Drosophila*, Piwi-family proteins are essential for male fertility. In contrast with mammals, *Drosophila* piRNAs contain a high proportion of sequences that are transposon-related. Brennecke *et al.* (2007) reports that *Drosophila* piRNAs regulates transposon activity in the germline by containing sequences that are able to recognise invasive parasitic genetic elements. Genetic memory of any invasive element is possible upon its transposition into an existing piRNA loci, leading to the subsequent silencing of related elements in the genome. In mammals, piRNAs have been shown to be involved in establishing DNA methylation imprints in the paternal germline (Daxinger and Whitelaw, 2012).

1.5.3 Small RNA-mediated gene activation

In contrast with the silencing activity of siRNAs and miRNAs, there is a class of small RNAs that induces the expression of its targets. In 2006, Li *et al.* (2006) observed a long-lasting and sequence-specific induction of E-cadherin, p21 and VEGF (vascular endothelial growth factor) in human cells after transfecting dsRNAs that targeted the promoter region of those genes. This effect was termed "RNAa" (RNA activation) by the authors of this paper.

Several parallels were discovered between this class of dsRNAs and miRNAs — firstly,

sequential mutation of the dsRNAs revealed that the sequence at the 5' end is critical for activity, similar to how the "seed" sequence is the main determinant of miRNA activity; secondly, Argonaute 2 is crucial for the function of these dsRNAs; and lastly, the effects of these dsRNAs were not caused by nonspecific interferon response.

A ChIP (chromatin immunoprecipitation) study on histone methylation revealed significant demethylation of lysine-9 in histone 3 (H3m2K9) in the E-cadherin promoter after the transfection of the activating dsRNA. As methylation of H3K9 is associated with repression of E-cadherin, the observed upregulation of E-cadherin is likely due to the demethylation of histone 3. The molecular mechanism that links RNAa to histone demethylation remains unclear (Li *et al.*, 2006).

Since the publication of the first paper regarding RNAa, several other papers have observed similar activatory effects for other dsRNAs on human cells. Janowski *et al.* (2007) reported the upregulation of progesterone receptor proteins upon transfection of dsRNAs targeting the promoter region upstream of the gene; similar upregulation was observed in VEGF-A (Turunen *et al.*, 2009) and LDLR (low-density lipoprotein receptor) (Matsui *et al.*, 2010).

The discovery of similar activatory dsRNA in two other primates and in mice indicates that RNAa is not a biological oddity that is specific to humans (Huang *et al.*, 2010b). Also, the source of dsRNAs could well be endogenous — it has been shown that there is at least three endogenous sources of the dsRNA involved in RNAa. Human miR-373 has been shown to directly upregulate the expression of CSDC2 (cold-shock domain-containing protein C2) (Li *et al.*, 2008a); human miR-369-3 upregulates the expression of TNFα (tumor necrosis factor-α) while let-7 upregulates HMGA2 (high-mobility group A2) (Vasudevan *et al.*, 2007). In the latter study, the target sites were located in the 3' UTR instead of the promoter region.

Although the molecular machinery underlying these observations remains unclear, the increasing evidence of small RNA involved in the upregulation of its targets has led to the coining of "saRNAs" (small activating RNAs) to describe this intriguing class of RNAs (Huang

et al., 2010b).

1.6 Aims of my project

In the first part of my work, I predicted the tissue-specific depletions of miRNAs based on the pattern of expression of their respective mRNA targets in *Drosophila*. The veracity of these predictions were tested experimentally in the wet lab. Details of the predictions made and experiment results are in Chapter 2.

Later on, as part of a larger collaboration, I studied the in-depth small RNAome of two marine organisms: *Stylophora pistillata*, a coral that inhabits many oceans of the world, and *Symbiodinium sp.*, one of the more abundant photosynthetic dinoflagellate that lives *in symbio* with the coral. Prior to the small RNA sequencing, I optimised a protocol that resulted in high quality RNA extracts from *Symbiodinium sp.* The RNA extraction protocol is detailed in Chapter 3, while the analysis of the small RNAome of both organisms is in Chapter 4.

Due to the nature of my projects, each chapter will contain a section dedicated to the discussion of results obtained in that chapter, instead of having a single chapter that discusses the results obtained from all the work that has been done.

Chapter 2

Identification of tissue-specific *Drosophila* miRNAs

2.1 Introduction

The previous chapter outlined the biogenesis of miRNAs, and the post-transcriptional down-regulation of their respective mRNA targets in the cell. Thus, the pattern of miRNA expression can be deduced based on the expression levels of its targets — if a tissue contains high levels of transcripts regulated by miR-1, it is likely that the expression of miR-1 is limited in that tissue. The tissue-specific pattern of miRNA expression can be computationally inferred from a comprehensive transcriptome that spans multiple fly tissues, and a dataset that describes transcripts targeted by fly miRNAs. The resulting predictions can then be verified in the wet lab, or compared to published data.

In this section, existing algorithms that predict miRNA-mRNA targeting are briefly outlined, followed by a review of methods that can be used to accurately assay miRNA levels in a tissue sample, and an overview of published tissue-specific measurements of miRNA expression. It explains the underlying factors behind my choice of MicroCosm Targets (Griffiths-Jones *et al.*, 2006) as the algorithm to predict miRNA-mRNA targeting that resulted in predicting the depletion of 10 miRNAs in a tissue-specific manner. In order to verify my predictions, the use of TaqMan RT-qPCR (real-time quantitative PCR) to assay for my miRNAs

of interest would be expounded in this section. These assays were performed on tissues that had to be dissected ourselves, as tissue-specific miRNA expression data from these tissues are not available in literature.

2.1.1 Algorithms predicting miRNA-mRNA targeting

Based on the principles of miRNA-mRNA targeting outlined in the previous chapter, many prediction algorithms have been written to predict the mRNA targets of miRNAs. A review of these programs can be found in Bartel (2009). As my work focuses on fly miRNAs, I will highlight algorithms that have been used to predict miRNA-mRNA binding in the fly. These four algorithms are MicroCosm Targets (Enright *et al.*, 2003; Griffiths-Jones *et al.*, 2006, 2008), TargetScanFly (Lewis *et al.*, 2003, 2005; Ruby *et al.*, 2007; Kheradpour *et al.*, 2007), PicTar (Grün *et al.*, 2005; Krek *et al.*, 2005; Lall *et al.*, 2006) and PITA (Kertesz *et al.*, 2007).

2.1.1.1 MicroCosm Targets (formerly miRBase Targets)

miRBase is a database with three main roles: the miRBase Registry is an independent arbiter of miRNA gene nomenclature, the miRBase Sequences is an online repository for all known miRNA sequence data and annotation, while miRBase Targets contains predictions of miRNA target genes (Griffiths-Jones *et al.*, 2006). The predictions in miRBase Targets uses an algorithm called miRanda, which was written a few years before miRBase Targets was available. miRanda predicted miRNA-mRNA binding based on three steps: sequence-matching to identify possible binding between two roughly complementary sequences (the short miRNA and the longer mRNA target binding site), determining the energetics of this physical interaction, and lastly filtering the results based on evolutionary conservation (Enright *et al.*, 2003).

The open-source miRanda algorithm has been continually refined and incorporated into a novel, fully automated pipeline called "MicroCosm", developed in-house by the Enright lab (shown in Figure 2.1). miRBase Targets incorporates code from miRanda and the most

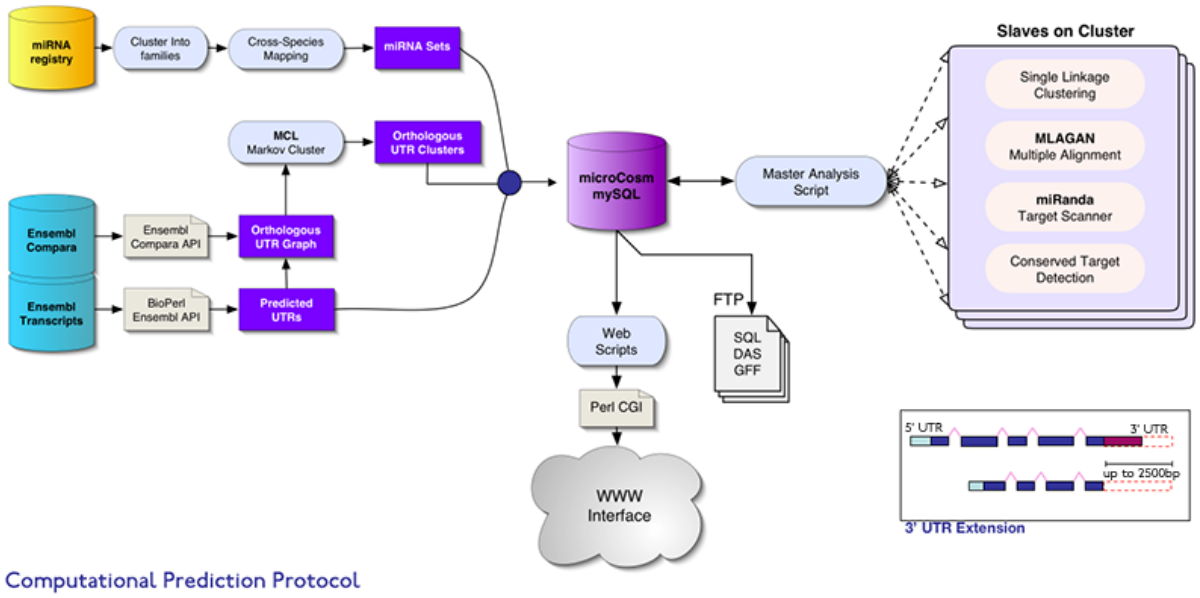


Figure 2.1: Overview of the MicroCosm Targets pipeline. Using the miRanda algorithm, MicroCosm Targets produces target predictions based on miRNA data from miRBase and UTR data from Ensembl. These predictions are available for download via FTP, while specific searches can be performed online as well. Diagram from `http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/info.html`.

up-to-date UTR data from Ensembl, and allows end-users to view the results from this analysis online (Griffiths-Jones *et al.*, 2006). The current version of the targets file is v5.0. As more validated miRNA-target sites were shown to have mismatches in the seed region, the pipeline lifted the constraint on having perfect seed matches as one of the criteria for prediction. Also, there have been reports about the importance of secondary structure in miRNA-target recognition, for example sequence accessibility, AU bias and position within the UTR, which the authors plan to consider in the next version of the software (Griffiths-Jones *et al.*, 2008).

2.1.1.2 TargetScanFly (TargetScanS implemented on fly data)

At the time TargetScan was initially written by Lewis *et al.* (2003), mRNA targets of known miRNAs had been identified in nematodes, insects and plants, but not in vertebrates. Unlike plants, target sites in animal 3' UTRs that were almost perfectly complementary to the miRNA sequences did not produce more *bona fide* hits than would be expected by chance —

in plants, a similar search produced hits with a signal-to-noise ratio exceeding 10:1.

Two years later, Lewis *et al.* (2005) presented a simplified version of the TargetScan algorithm, known as TargetScanS. Multiple criteria taken into consideration in the predecessor, such as the thermodynamic stability of miRNA-mRNA pairing outside the immediate vicinity of the miRNA seed sequence, or presence of multiple complementary sites per UTR were all discarded. With the increase in genomic sequences available from more vertebrates, the use of conservation and the consideration of the primary sequence of the target site were sufficient to determine the likelihood of a miRNA binding to the target site.

Another two years later, the use of new computational methods that predict miRNA sequences from small RNAseq efforts resulted in the increase of known miRNAs to 148 miRNAs (Ruby *et al.*, 2007). TargetScanS utilised this database of 148 miRNAs and UTR data from 12 sequenced *Drosophila* species to produce a set of miRNA-mRNA target predictions in Drosophilids. Similar to previous observations, the requirement of perfect conservation across all 12 species maximises the confidence in the predicted targets, at the expense of a substantial reduction in the sensitivity of the algorithm.

To overcome the drop in sensitivity, a branch length score (BLS) was calculated for target sites which were not completely conserved across all 12 species. This score ranges from 0 (not conserved at all) to 1 (completely conserved), and corresponds to the distribution of species where the target site was present — for example, a site present in only two species would have a higher BLS score if the two species were more distantly related to each other, than if the two species shared a very recent common ancestor. (Kheradpour *et al.*, 2007) showed that a BLS of 0.60 produced more matches to experimentally verified targets than BLS of 1 (complete conservation). These predictions are available online as TargetScanFly (`http://www.targetscan.org/fly/`).

2.1.1.3 PicTar

Previous computational efforts at identifying miRNA targets can identify targets for single miRNAs, but have not been used to score common targets of several miRNAs, which might

be coexpressed in specific tissues or developmental stages. Also, those methods tend to produce high false-positive rates when the number of binding sites on a given 3' UTR is small. PicTar (probabilistic identification of combinations of target sites) overcomes these problems by generalising previous methods, allowing for the identification of targets for both single miRNAs and combination of miRNAs (Grün *et al.*, 2005; Krek *et al.*, 2005; Lall *et al.*, 2006).

When applied to flies, similar to the work done in vertebrates by Krek *et al.* (2005), PicTar used the data from seven sequenced *Drosophila* species to predict and analyse miRNA target in flies. As the inclusion of evolutionary conservation in the prediction algorithm improved the signal-to-noise ratio in identifying *bona fide* miRNA-mRNA interactions considerably, the whole genome sequence of seven Drosophilids (*D. melanogaster*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis*) was used for the PicTar analysis.

The output from PicTar was then compared to published *Drosophila* miRNA-mRNA interactions published in literature. The results are summarised in Table 2.1. With setting S1, PicTar managed to recover 8/9 of all known targets with experimental *in vivo* evidence and 4/10 of targets with other experimental support (Grün *et al.*, 2005).

2.1.1.4 PITA

Contrary to most other prediction programs that evaluates the closeness of the sequence match between the microRNA and the target mRNA, PITA (Probability of Interaction by Target Accessibility) assesses the secondary structure in which the target is embedded and calculates the energetic cost needed to make the target accessible for miRNA binding. There have been studies considering the effect of secondary structure on miRNA binding, but the effect had never been experimentally quantified, nor had it been applied in a computer model that could produce a genome-wide prediction of miRNA targets. A brief schematic of this algorithm is shown in Figure 2.2.

To assess the predictive power of PITA relative to PicTar or miRanda, the authors com-

Category	microRNA–Target	S1	S2	S3	Comments
microRNA targets with experimental support [4,14,24]	*bantam–hid*	+	+	+	
	miR-7–hairy	+	–	–	Not strictly conserved in all flies but scattered sites present
	miR-7–HLHm3	+	+	+	
	miR-7–m4	+	+	+	3′ UTR absent in FlyBase 4.1 annotation
	miR-4–Bearded	+	–	–	Not conserved in all flies
	miR-4–bagpipe	+	+	+	
	miR-2–sickle	+	+	+	
	miR-2–reaper	+	–	–	Not conserved in all flies
	miR-2–grim	–	–	–	Nucleus consists of six Watson–Crick basepairings and one G/U
microRNA targets with experimental support [12] (Luciferase reporter assays in cell lines)	*bantam–MAD*	–	–	–	
	miR-287–CRMP	–	–	–	
	miR-7–HLHm5	+	+	+	
	miR-279–SP555	+	+	+	
	miR-310–imd	+	+	+	Recovered if *miR-310* presumed to be conserved in all flies
	miR-1–tutl	–	–	–	
	miR-34–su(z) 12	–	–	–	Not recovered because nucleus overlaps with repeat
	miR-12–rt	–	–	–	
	miR-124–gli	+	+	+	
	miR-7–fng	–	–	–	
False positives according to experiments [12]	*miR-287–dip1*	–	–	–	
	miR-303–CG14991	–	–	–	
	miR-278–tup	–	–	–	
	miR-317–yellow-c	–	–	–	
	miR-318–CG13380	–	–	–	
	miR-286–boss	+	+	+	
	miR-288–CG32057	–	–	–	
	miR-276b–ke1	–	–	–	
	miR316–ia2	–	–	–	

Table 2.1: Summary of comparison between *in silico* PicTar predictions and *in vivo* experimental work. S1, a high sensitivity setting, required anchor conservation across *D. melanogaster*, *D. yakuba*, *D. ananassae* and *D. pseudoobscura*, and no free energy filtering of the anchor sites; S2 used unmasked repeats but anchors had to be conserved in all flies; S3 was equivalent to S1 but required conservation of anchors in all flies (table adapted from Grün *et al.* (2005)).

piled a database of 190 experimentally-tested miRNA-mRNA interactions, and this had a binary classification — either they did, or did not interact. PITA, without the use of filters such as conservation or other statistical criteria employed by the other algorithms, achieved a slightly better sensitivity and specificity than both PicTar and miRanda (Kertesz *et al.*, 2007).

2.1.2 Existing methods of assaying miRNAs

Detection and quantification of miRNAs is mainly achieved through the use of small RNA sequencing, microarrays, real-time PCR and Northern blots. A comparison of the more popular methods used to profile miRNA expression is shown in Table 2.2.

For my experiments, I opted to use RT-qPCR to assay my miRNAs of interest. Among

MicroRNA-profiling technologies

Advantages	Disadvantages	Assay or platform*	Vendor	RNA required	Material costs per sample‡
Quantitative reverse transcription PCR (qRT-PCR)					
Established method, sensitive and specific. Can be used for absolute quantification	Cannot identify novel microRNAs (miRNAs) (which is problematic for less well-studied organisms in which the miRNA repertoire is not well-defined). Only medium-throughput with respect to the number of samples processed per day	TaqMan individual assays	ABI	<ng or ng–µg	$$
		miRCURY LNA qPCR	Exiqon		
		TaqMan OpenArray	ABI		
		TaqMan TLDA microfluidics card	ABI		
		Biomark HD system	Fluidigm		
		SmartChip human microRNA	Wafergen		
		miScript miRNA PCR array	SABiosciences/ Qiagen		
MicroRNA microarray					
Established method. Fairly low-cost and high-throughput with respect to the number of samples that can be processed per day	Typically lower specificity than qRT-PCR or RNA sequencing. Difficult to use for absolute quantification. Cannot typically identify novel miRNAs	Geniom Biochip miRNA	CBC (febit)	ng–µg	$
		GeneChip miRNA array	Affymetrix		
		GenoExplorer	Genosensor		
		MicroRNA microarray	Agilent		
		miRCURY LNA microRNA array	Exiqon		
		NCode miRNA array	Invitrogen		
		nCounter (not a microarray but hybridization-based)	Nanostring		
		OneArray	Phalanx Biotech		
		Sentrix array matrix and BeadChips	Illumina		
		µParaFlo biochip array	LC Biosciences		
RNA sequencing: high-throughput next-generation sequencing platforms					
High accuracy in distinguishing miRNAs that are very similar in sequence, as well as isomiRs. Can detect novel miRNAs	Substantial computational support needed for data analysis. Cannot be used for absolute quantification	HiSeq 2000 (or Genome Analyzer IIX)	Illumina	ng–µg or >µg	$$$
		SOLiD	ABI		
		GS FLX+ (454 sequencing)	Roche		
RNA sequencing: smaller-scale next-generation sequencing platforms					
High accuracy in distinguishing miRNAs that are very similar in sequence, as well as isomiRs. Can detect novel miRNAs	Substantial computational support needed for data analysis. Cannot be used for absolute quantification	Ion Torrent	Invitrogen	ng–µg or >µg	$$$
		MiSeq	Illumina		
		GS Junior (454)	Roche		
RNA sequencing: single-molecule sequencing technologies					
Amplification not required. Potential to determine absolute quantification	Expensive, not widely accessible, single-molecule real-time approach not yet demonstrated for miRNAs	tSMS	Helicos	ng–µg	Not yet defined for miRNA-seq
		SMRT	Pacific Biosciences		

*This is not meant as a comprehensive list, but as a sample of commercially available platforms. ‡Excludes cost of instrument; the more '$' symbols there are, the more expensive the materials per sample are.

Table 2.2: A comparison of the three most popular miRNA profiling technologies in common use today: RT-qPCR, microarrays and RNA sequencing. Compared to sequencing and microarrays, RT-qPCR techniques require the least amount of initial RNA, and strike a middle ground in terms of cost and specificity. Although the throughput of RT-qPCR is lowest among these three techniques, the results produced from qPCR experiments are easy to interpret. Table obtained from Pritchard *et al.* (2012).

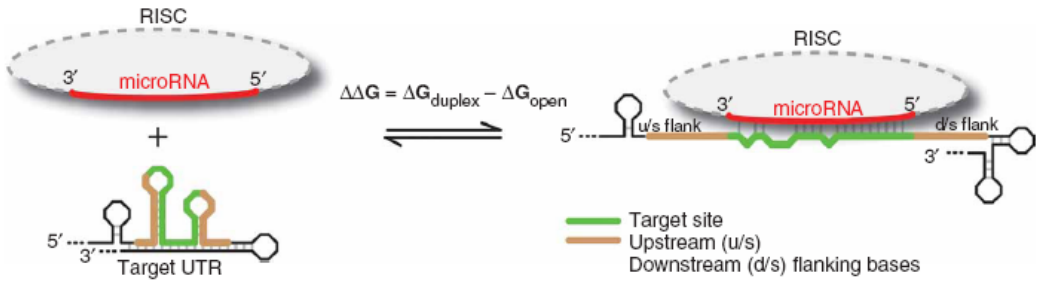


Figure 2.2: Pictorial representation of ΔΔG, the energy required or produced by the miRNA-mRNA interaction. ΔG_{duplex} refers to the energy produced by the binding of miRNA to mRNA, while ΔG_{open} refers to the energy required to make the target accessible for binding. The more exergonic ΔΔG is, the higher the likelihood for binding. Figure adapted from Kertesz *et al.* (2007).

the PCR-based techniques available, TaqMan was chosen as it is a sensitive, fluorophore-based method to assay DNA concentrations. The probe used in TaqMan experiments is a dual-labelled probe containing a fluorophore at the 5' end and a quencher at the 3' end. During the annealing step, the probe binds to the amplicon at a site downstream of the PCR primer binding site. The assay takes advantage of the 5'–3' exonuclease function of Taq polymerase to degrade the probe during the extension step of the PCR reaction, in a manner reminiscent of the classic PacMan video game character (which inspired the "-Man" suffix of the TaqMan name). When the probe is degraded, the fluorophore is separated from its quencher, allowing the accurate quantification of the amplicon by detecting the accumulation of fluorescence (Heid *et al.*, 1996; Leutenegger, 2001; Walker, 2002). Figure 2.3 gives an overview of the mechanism of the TaqMan assay.

A few years after its invention in 1996, the TaqMan method had been successfully modified to quantify longer mRNAs by inserting a reverse transcription step before the assay (Wang and Brown, 1999). This has been adapted to quantify smaller miRNAs by using stem-loop primers in the reverse transcription step (Chen *et al.*, 2005). The use of these stem-loop primers with a short ∼6 bp overhang (as seen in Figure 2.4) has four advantages. One, it increases the specificity of the primer to the targeted mature miRNA; two, it prevents the primer from binding to longer RNAs containing sequence complementary to the overhang; three, the stem-loop structure increases the melting temperature of the primer-

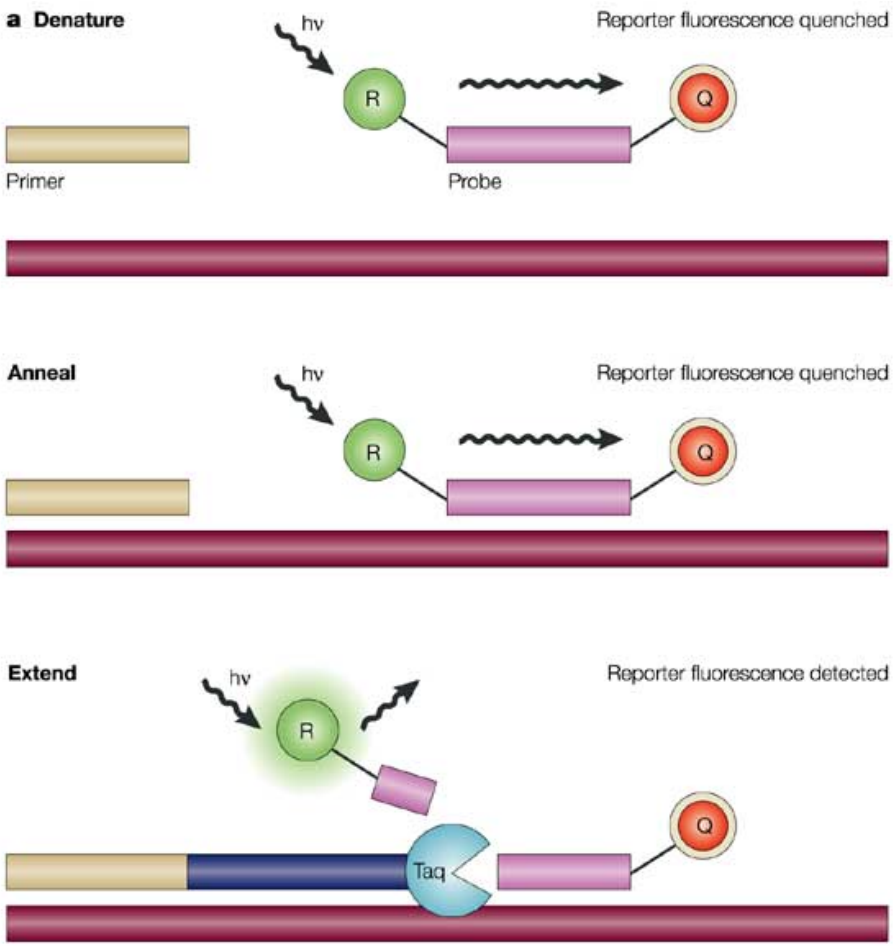


Figure 2.3: Overview of TaqMan probe hydrolysis during qPCR amplification. During the extension step of PCR, the 5'–3' exonuclease function of Taq hydrolyses the TaqMan probe, separating the fluorophore from its quencher. This results in fluorescence being produced proportional to the concentration of the product being formed. Figure from Koch (2004).

DNA heteroduplex during reverse transcription, due to base stacking at the stem, resulting in increased efficiency of the reaction; and four, as the stem-loop unfolds in the qPCR step, it provides a longer template more suited for TaqMan assays (Schmittgen *et al.*, 2008).

While TaqMan is cheaper than microarrays or sequencing techniques, the TaqMan probe can be quite costly, especially when different fluorescent probes are used for each miRNA of interest. Some researchers circumvent this issue by using generic TaqMan probes that anneal to standardised sequences introduced either through the forward primer (Whitcombe *et al.*, 1998) or the stem loop primer used in reverse transcription (Varkonyi-Gasic *et al.*, 2007). The former is more suitable for assays of longer mRNAs, while the latter is more suitable for the ∼22 nt miRNAs which I was interested in assaying. Although the fluorescent probe used in the Varkonyi-Gasic *et al.* (2007) protocol is only of 8 nt in length, the inclusion of a LNA (locked nucleic acid) base in the probe ensures specificity as LNA bases have a much higher hybridisation affinity to perfectly matched RNA or DNA than to sequences with any number of mismatches (Vester and Wengels, 2004).

My miRNA assays broadly follow the method described in Varkonyi-Gasic *et al.* (2007), which presented a cost-efficient method of assaying miRNA in plants. This method has never been shown to work with animal miRNAs, but my experiments have demonstrated high specificity and accuracy on extracted fly tissues, which is illustrated in Figure 2.4.

2.1.3 Previous spatiotemporal surveys of *Drosophila* miRNA expression

Prior to the widespread use of RNAseq in quantifying small RNA expression, large scale profiling of miRNA expression across developmental stages or dissected tissues relied on cloning and sequencing. One such study, carried out by Aravin *et al.* (2003), sequenced 382 miRNA clones across nine developmental time points as well as in adult testes. As the expression of many miRNAs are relatively low, the majority of the miRNAs had less than 20 clones spread across the ten libraries under study — as such, it is hard to draw quantitative comparisons between time points or tissues based on the clone counts.

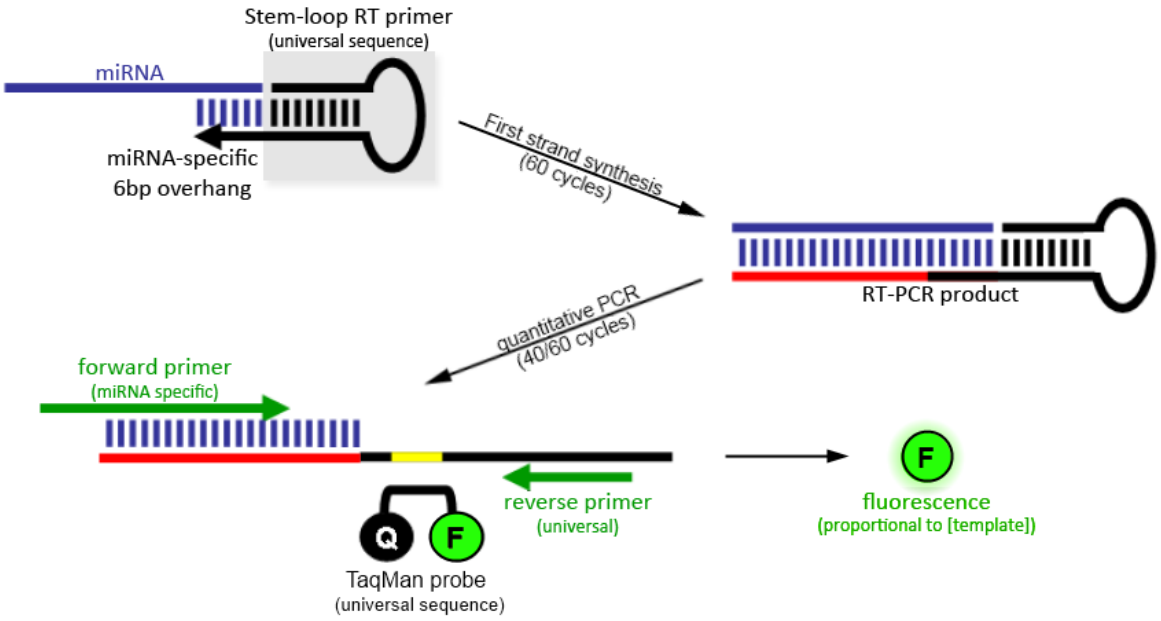


Figure 2.4: Overview of my miRNA assay. Specificity for the target sequence is achieved by the use of the miRNA-specific 6 bp overhang in the initial RT primer, as well as the miRNA-specific forward primer in the qPCR. As the forward primers cover the remaining 15–18 nt of the miRNA sequence, a CG-rich 6 bp over hang is added to the forward primer to increase the melting temperature of the primer. The region marked in yellow is the 8 bp binding site for the TaqMan probe, which is hydrolysed during extension phase. The fluorescence produced is thus proportional to the concentration of the template. Figure adapted from Varkonyi-Gasic *et al.* (2007), and edited for clarity.

In the same study, Aravin *et al.* (2003) carried out a Northern blot analysis on a smaller subset of miRNAs to verify the coexpression of miRNAs based on the genomic locations of the miRNA precursors. Coexpressed miRNAs were predicted to have very similar expression profiles across the ten libraries. While Northern blots are sufficient to provide qualitative comparisons across developmental time points or tissues, it is not as sensitive or accurate as PCR-based or sequencing methods in assaying miRNAs.

The modENCODE project (**Mod**el Organism **ENC**yclopedia **O**f **D**NA **E**lements, `http://modencode.org`), which started in 2007, is an ongoing project that aims to identify all functional elements in the *D. melanogaster* and *C. elegans* genomes (Celniker *et al.*, 2009). Efforts to study the expression of small RNAs across numerous cell lines, developmental time points and tissues has, as of now, produced RNAseq data from 80 different *D. melanogaster* libraries. Some of these published libraries can be found in Chung *et al.* (2008); Flynt

et al. (2009); Okamura *et al.* (2009, 2011); Berezikov *et al.* (2011). Outside of modENCODE, there are hundreds of other libraries — for instance, the work reported in Berezikov *et al.* (2011) was carried out on sequence data from 191 datasets, which had a mix of both modENCODE and non-modENCODE data.

Despite the huge range of data available for *D. melanogaster*, the tissue-specific datasets available center mostly around adult fly heads, ovaries and testes due to the ease of mass dissections. There are no libraries that specifically focus on adult fly tissues such as Malpighian tubules, gut or crop. Also, while there are data from fly embryos and larva at different developmental time points, there is no sequence data that is specific for larval tissues.

As the verification of my computational predictions required data from tissues that were not available in the literature, these tissues had to be dissected prior to measuring the miRNA expression levels via RT-qPCR assays.

2.2 Materials and methods

2.2.1 Algorithm of choice for miRNA-mRNA targeting

In my analysis, I have chosen MicroCosm Targets as the basis of my own predictions. MicroCosm Targets strikes a good balance of considering factors that have experimental support while being open to other factors that might prove to be essential with more experimental support. Examples of factors with experimental support include using seed sequences as the primary basis of target prediction (in a manner that is less strict than that in TargetScanFly), filtering target sites based on conservation to lower false positives, and considering the free energy of the miRNA-mRNA duplex (but not as reliant on it as PicTar); while an example of potentially important factors include secondary features for miRNA-mRNA recognition (on which PITA is based completely). MicroCosm Targets is actively updated and maintained over the years by the Enright Lab at EMBL, but the same can't be said about PicTar or PITA.

For PITA, there has been a paper disputing the claims that the algorithm achieves better

sensitivity and specificity than other available programs. Chen *et al.* (2009) noted that PITA — which does not filter predictions based on evolutionary conservation — achieves lower specificity at predicting miRNA-mRNA interaction than other algorithms that employ the filter.

The primary drawback of using miRBase is that it has not incorporated the additional fly miRNAs discovered by Ruby *et al.* (2007), but it is expected that the inclusion of additional miRNA sequences will not drastically change the list of miRNAs that my work has identified as being interesting.

2.2.2 Databases used

miRNA target prediction data were obtained from miRBase Targets v5 (`http://microrna.sanger.ac.uk/targets/v5/`) (Enright *et al.*, 2003; Griffiths-Jones *et al.*, 2006, 2008). The algorithm scanned 93 known miRNA sequences and 15,416 *Drosophila* genes, producing 38,772 predicted miRNA-target interactions. On average, every miRNA is predicted to target about 417 mRNA transcripts, while every mRNA is targeted by about 2.5 miRNAs.

For *Drosophila* tissue-specific gene expression data, the FlyAtlas dataset (`http://www.flyatlas.org`) was used. FlyAtlas is a database of gene expression data generated with the Affymetrix microarray platform from multiple *D. melanogaster* adult tissues. In its first data release, it contained expression data for 18,500 transcripts for each of 13 tissues (11 tissues from the adult fly: adult carcass, brain, crop, head, hindgut, male accessory glands, midgut, ovary, testis, thoracicoabdominal ganglion and tubules, and 2 from larval tissues: fat body and tubules). The expression levels of the transcript in each tissue relative to the whole fly are denoted by three possible "AffyCalls": "UP", "NONE" or "DOWN". "UP" means that there is significant upregulation of gene expression in the tissue, "NONE" denotes gene expression in the tissue is not significantly different from that of the whole fly, and "DOWN" translates to the significant downregulation of gene expression in the tissue. These AffyCalls are assigned by the proprietary GCOS v1.4 software from Affymetrix (Chintapalli *et al.*, 2007).

Work on FlyAtlas did not stop after publication, as more tissues have been added to the database. However, as the data updates occurred after my experimental work commenced, my experiments have relied on predictions based on the older FlyAtlas data. Based on the updated predictions of Gorton and Micklem (2009) (unpublished), which repeated my methodology on the updated FlyAtlas data (containing five extra tissues), two extra miRNAs were added to the initial assays (see Section 2.3.1).

2.2.3 RNA extraction from fly tissues

All fly tissue RNA were extracted using guanidium thiocyanate-phenol-chloroform (marketed as TRIzol, Invitrogen) (Chomczynski and Sacchi, 1987).

Briefly, fly tissues were homogenised in TRIzol on ice, followed by centrifugation. The pellet was discarded, and chloroform was added to the supernatant. The mixture was left to stand for a few minutes, followed by centrifugation. The upper phase was removed to a fresh tube, and isopropanol was added to precipitate the RNA. After an hour of incubation, the mixture was centrifuged. The supernatant was then discarded, while the pellet was resuspended in DEPC water and kept. To check the purity of the extracted RNA, a NanoDrop machine was used. If the 260/230 ratio of any sample was less than 1.8, the sample was then purified via ethanol precipitation overnight to remove contaminants.

The number of tissues dissected for the experiments were based on estimates needed to extract about 50 µg of total RNA from larger tissues (head, ovary, testis, whole fly) or about 5 µg of total RNA from smaller tissues (all other tissues). I carried out dissections for the larger tissues, while Venkat Chintapalli, a collaborator from the Dow Lab (University of Glasgow), carried out dissections for the smaller tissues.

For each tissue, the approximate number of dissected tissues is listed in Table 2.3. All of the adult tissues were dissected from week-old flies, while the larval tissues were dissected from third-instar larva. Except for the dissections of testes and ovaries, equal numbers of males and females were used for the dissection of all tissues. Canton S flies — from the same fly stock that was used in FlyAtlas — were used to ensure consistency across all dissected

tissues from both labs, avoiding the potential pitfall of using fly stocks that had slightly different transcriptomic landscapes than that described in FlyAtlas.

2.2.4 TaqMan RT-qPCR miRNA assay

2.2.4.1 List of small RNAs and miRNAs assayed

Based on computational predictions from Liew and Micklem (2008); Gorton and Micklem (2009) (discussed in Section 2.3.1), 10 miRNAs were assayed. These miRNAs are: let-7, miR-1, miR-2a, miR-11, miR-79, mir-92a, miR-92b, miR-277, miR-1013 and miR-iab-4-3p. There were two positive controls in my assays — one of known concentration, independent of the extracted RNA, and another of concentration proportional to the miRNAs in the extracted RNA. For the former, a synthetic 26-mer RNA ("baurb-000002", stock concentration 0.1 µg/µl) of sequence 5'-GUAUCUCACGUGAUACCAGCGAUUCC-3' was used; for the latter, 2S ribosomal RNA (rRNA), the smallest of all rRNAs present in flies, was used. It was chosen due to its size of 30 bp, which is similar to that of the miRNAs assayed (20–24 bp), and for its ubiquitous expression across all fly tissues.

Each miRNA assayed had two primers specific to its sequence: the stem-loop primer used in the initial RT-PCR, and the forward primer used in the subsequent qPCR (see 2.4). The 3' end of the stem-loop primer has a 6 bp overhang specific to the 3' end of the miRNA assayed; the forward primer would then cover the remaining length of the miRNA (i.e. if the miRNA is 22 bp, the forward primer binds to 16 bp of the miRNA from the 5' end). The 5' end of the forward primer has a GC-rich 6 bp 5' extension as well, in order to increase its melting temperature, especially important if the miRNA is AT-rich (Chen *et al.*, 2005; Varkonyi-Gasic *et al.*, 2007). Primers were designed and their suitability assessed using PerlPrimer (Marshall, 2004). Table 2.4 lists the sequences of these miRNAs, as well as the specific primers that target these miRNAs.

Tissue	Definition (from FlyAtlas.org)	Amount per replicate (approx)	Dissected by
Adult tubule	Both anterior and posterior tubules with their common ureters, severed at the junction with the gut	30	VC
Adult midgut	From (and including) the proventriculus, down to just in front of the insertion of the Malpighian tubules	20	VC
Adult hindgut	From the insertion of the tubules, back to and including the rectum	30	VC
Adult crop	The round diverticulum of the foregut, including the stalk	20	VC
Adult male accessory gland	Accessory glands excluding other parts of the male genital tract	20	VC
Larval fat body	Prominent lateral fat bodies	10	VC
Larval tubule	Both anterior and posterior tubules with their common ureters, severed at the junction with the gut	30	VC
Larval midgut	From (and including) the proventriculus, down to just in front of the insertion of the Malpighian tubules	20	VC
Larval hindgut	From the insertion of the tubules, back to and including the rectum	30	VC
Adult head	Severed at the neck. Includes brain, eyes, cuticle and some fat body	120	YJL
Adult ovary	Ovaries from mated females, excluding the uterus and spermatheca	50	YJL
Adult testis	Testis excluding the accessory glands	100	YJL
Whole fly	Entire fly	30	YJL

Table 2.3: List of dissected tissues with number of tissues dissected, per replicate, in triplicate. Definitions of the dissected tissues are taken from `http://www.flyatlas.org`. The dissections of individual tissues were carried out, as noted, by Venkat Chintapalli ("VC") and me ("YJL").

Template	Sequence	Length	RT primer, 50 bp		qPCR Forward primer sequence			
			Universal 5' sequence, 44 bp	Specific, 6 bp	Random	Specific, (template - 6) bp	GC%	T_m/ °C
2S rRNA	UGCUUGGACUACAUAUGGUUGAGGGUUGUA	30	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	TACAAC	GCCCGC	TGCTTGGACTACATATGGTTGAGG	56.7%	73.76
baurb-000002	GUAUCUCACGUGAUACCAGCGAUUCC	26	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	GGAATC	CCAGCC	GTATCTCACGTGATACCAGC	57.7%	70.61
let-7	UGAGGUAGUAGGUUGUAUAGU	21	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	ACTATA	GGCGGG	TGAGGTAGTAGGTTG	61.9%	66.65
miR-1	UGGAAUGUAAAGAAGUAUGGAG	22	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	CTCCAT	GGCGGG	TGGAATGTAAAGAAGT	50.0%	65.21
miR-1013	AUAAAAGUAUGCCGAACUCG	20	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	CGAGTT	CCGCCG	ATAAAAGTATGCCG	55.0%	63.98
miR-11	CAUCACAGUCUGAGUUCUUGC	21	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	GCAAGA	CGCCCG	CATCACAGTCTGAGT	61.9%	69.45
miR-277	UAAAUGCACUAUCUGGUACGACA	23	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	TGTCGT	CGCCCG	TAAATGCACTATCTGGT	52.2%	67.29
miR-2a	UAUCACAGCCAGCUUUGAUGAGC	23	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	GCTCAT	GCCGCC	TATCACAGCCAGCTTTG	60.9%	70.66
miR-79	UAAAGCUAGAUUACCAAAGCAU	22	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	ATGCTT	GCCGCC	TAAAGCTAGATTACCA	50.0%	64.53
miR-92a	CAUUGCACUUGUCCCGGCCUAU	22	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	ATAGGC	CCGTTC	CATTGCACTTGTCCCG	59.1%	68.73
miR-92b	AAUUGCACUAGUCCCGGCCUGC	22	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	GCAGGC	CCGCCA	AATTGCACTAGTCCCG	59.1%	68.55
miR-iab-4-3p	CGGUAUACCUUCAGUAUACGUAAC	24	GTTGGCTCTGGTGCAGGGTCCG-AGGTATTCGCACCAGAGCCAAC	GTTACG	GGGCGG	CGGTATACCTTCAGTATA	54.2%	67.75

Table 2.4: Sequences of templates and primers used for the RT-qPCR assays. The random 6 bp sequences for the forward primer increases the primer's melting temperature. Care was taken to select sequences that were unlikely to form primer dimers with itself or with the reverse primer. The GC% and melting temperature (T_m) of the qPCR forward primer were calculated by PerlPrimer (Marshall, 2004).

2.2.4.2 Sensitivity, accuracy and specificity of assay

A set of experiments were carried out to assess the sensitivity and accuracy of the assay under controlled conditions. It was confirmed that my assay was able to pick up single-molecule levels of "baurb-000002" — the synthetic 26-mer RNA used as a positive control — in the absence of background RNA (see Figure 2.5), demonstrating the high sensitivity of the assay. Although the assay is able to pick up very low levels of RNA, those readings are less accurate as the assay works intermittently at such levels.

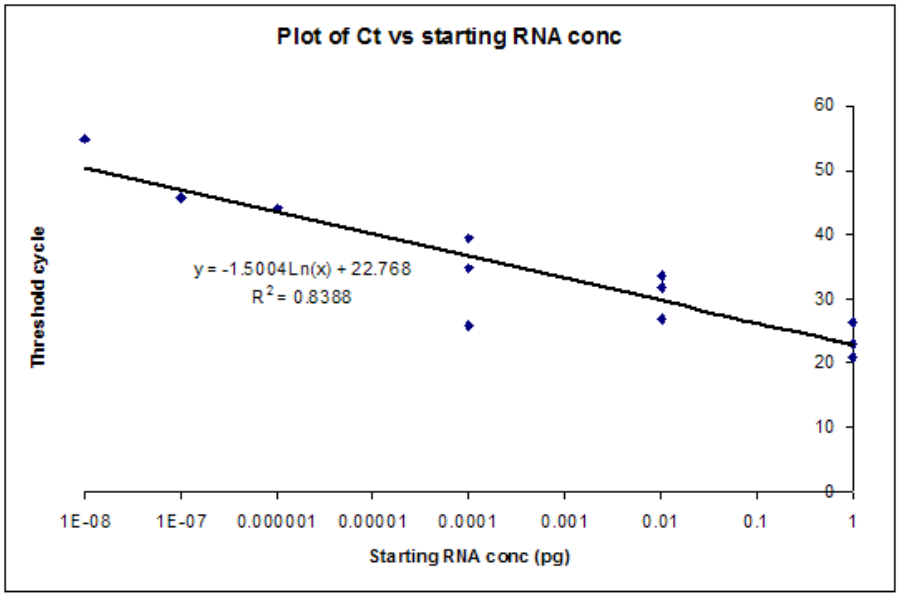


Figure 2.5: Experiment indicating single molecule sensitivity of assay in the absence of background RNA. This experiment assayed for the known RNA at six different concentrations, and for each concentration, three readings were taken (technical triplicate). Calculations indicate that the concentration of 10^{-8} pg of "baurb-000002", the synthetic 26-mer RNA, corresponds to about one molecule of that RNA. For the one, ten and hundred molecule level of the RNA, the assays fail twice out of three times. Another experiment (Table 2.5) shows that the known RNA is detected all three times at the 1,000 molecule level in the absence of background RNA.

To assess the robustness of the assay under varying total RNA concentration, equal and known concentrations of "baurb-000002" (5 pg) were assayed in varying backgrounds of total RNA extracted from whole flies. This experiment did not test the accuracy of the assay at very low RNA levels, as it was thought then that most miRNAs would not be present in such low quantities. The assay was highly accurate over five orders of background RNA (see Figure 2.6).

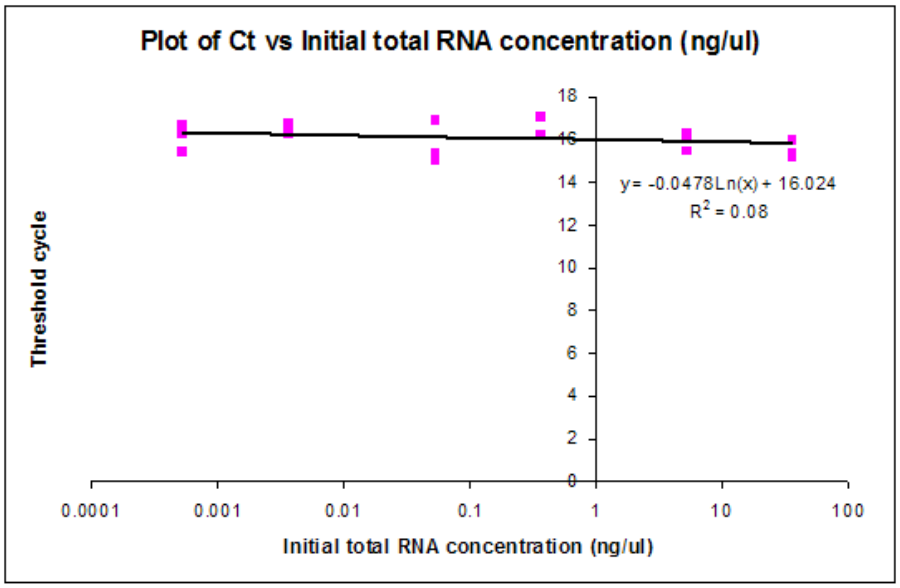


Figure 2.6: Experiment indicating accuracy over five orders of background RNA. Known amounts of RNA (5 pg) was probed in background RNAs of amounts ranging from 0.27 pg to 18 ng.

However, after carrying out assays on real miRNAs, it was realised that individual miRNAs are of concentrations about 6–10 orders of magnitude smaller than that of the total RNA. Another experiment was then designed to assess the assay performance when the probed RNA is of 2–11 orders of magnitude smaller in amount than the known background RNA. The results of the assay is shown in Table 2.5, Figure 2.7 and Figure 2.8.

Several conclusions can be drawn from this experiment. Regarding sensitivity, it is shown that the assay is able to detect RNA reliably when it is present above the 10,000 molecule level, despite having background RNA of up to 10 orders of magnitude more abundant than the assayed RNA. The assay can also detect RNA present below the 1,000 molecule threshold, but once the background RNA exceeds the assayed RNA by 7 orders of magnitude or more, the readings are inaccurate if it works at all. Regarding accuracy, there are two major conclusions: one, the accuracy of the assay suffers when the probed RNA is more than 7 orders of magnitude less abundant in amount than the background RNA (see Figure 2.8); and two, that probing of RNA above the 10,000 molecule level can be considered accurate, as seen from the raw data of the experiment (Table 2.5). The assay reliably picks up the assayed RNA when there are at least ∼7,500 molecules present, and the variance of these readings

	732	7320	73200	732000	73200000	molecules assayed
	0.00001	0.0001	0.001	0.01	1	RNA (pg)
0	43.14	33.88	30.65	28.19	23.97	
	37.92	33.08	30.96	28.02	25.01	
	36.79	34.9	31.85	27.84	25.03	
	39.28	33.95	31.15	28.02	24.67	
	3.86	0.95	0.70	0.18	0.70	
100	39.01	34.9	31.35	27.55	21.63	
	49.19	35.36	31.04	27.94	19.37	
	40.71	34.5	31.61	28.17	18.87	
	42.97	34.92	31.33	27.89	19.96	
	6.22	0.44	0.29	0.34	1.67	
1000	ND	34.95	31.43	28.15	17.88	
	ND	35.71	31.07	27.91	20.02	
	37.56	35.84	30.82	28.57	25.02	
	37.56	35.50	31.11	28.21	20.97	
	N/A	0.55	0.32	0.36	4.05	
10000	36.64	34.87	31.95	26.83	19.57	
	39.33	35.25	32.53	27.44	20.01	
	ND	34.52	31.66	28.92	19.52	
	37.99	34.88	32.05	27.73	19.70	
	1.35	0.37	0.48	1.19	0.31	
100000	41.94	39.28	34.88	28.16	20.63	
	44.23	42.02	1.42	32.51	19.74	
	41.9	35.97	33.51	31.64	20.58	
	42.69	39.09	34.20	30.77	20.32	
	1.54	3.12	0.69	2.61	0.58	
1000000	ND	41.84	36.29	33.28	20	
	12.39	38	32.28	33.07	19.76	
	ND	35.84	44.91	31.51	37.67	
	N/A	38.56	37.83	32.62	19.88	
	N/A	3.28	7.08	1.11	0.12	
Total RNA (pg)	40.10	36.15	32.94	29.21	20.92	

Legend

: average(), excluding anomalous readings.

: average of all blue boxes.

: max difference from mean, excluding anomalous readings.

red text : anomalous.

Note: ND = not detected

N/A = not applicable

Table 2.5: Raw data from the sensitivity/accuracy assay. The number of molecules of the assayed RNA is calculated to three significant figures. There is a general increase of threshold cycle (C_t) values with increasing background RNA. The data here is further analysed in Figure 2.7 and Figure 2.8.

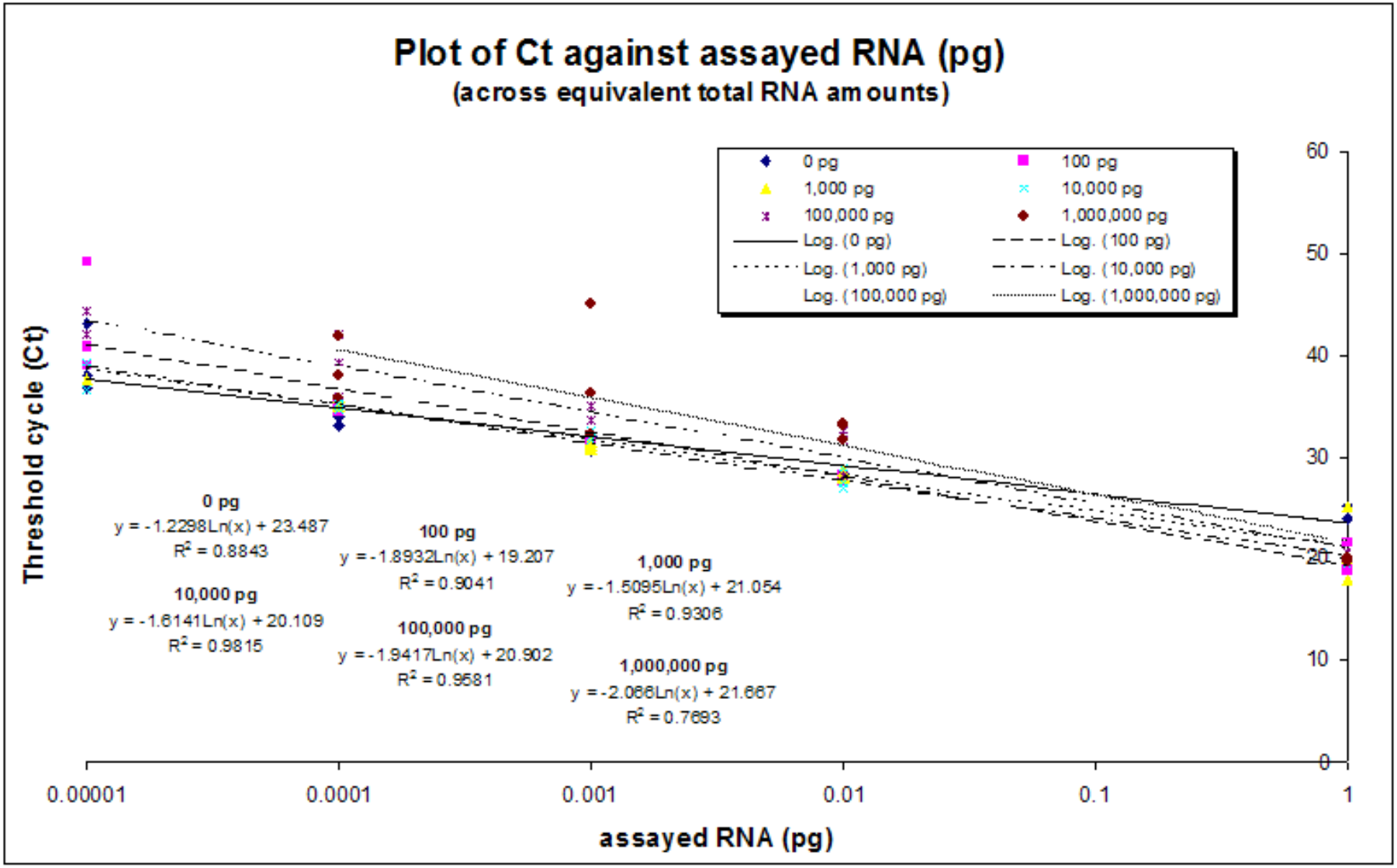


Figure 2.7: Graph of the sensitivity/accuracy assay, with trendlines drawn for different background RNA amounts. Note that the gradients are about the same (with the exception of the "0 pg" line).

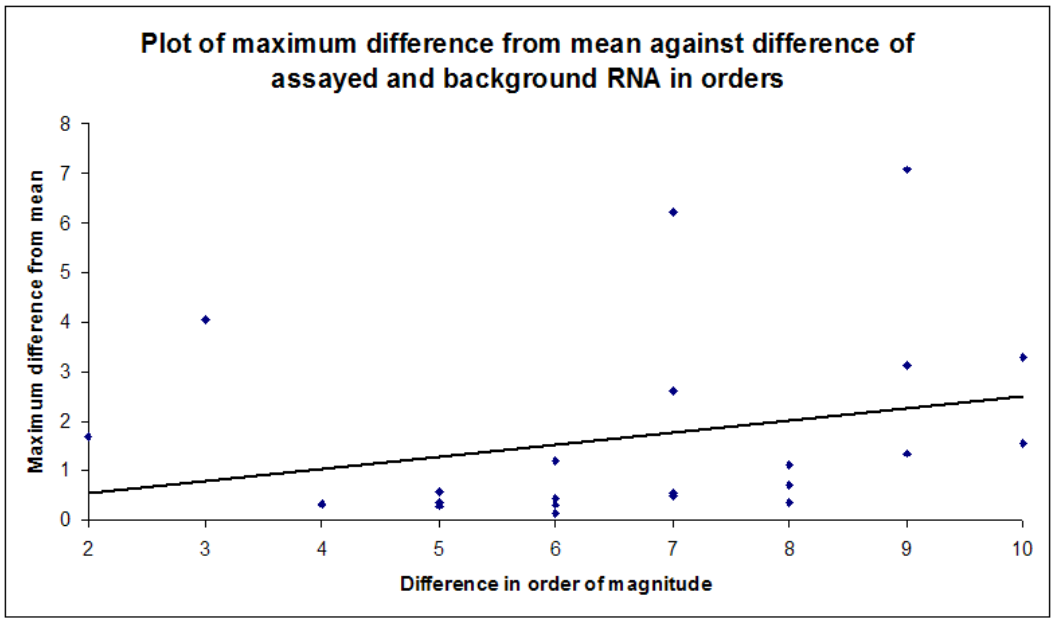


Figure 2.8: Graph of errors (orange boxes in Table 2.5) against difference of assayed and background RNA, by order of magnitude. Variances steadily increase with increasing order, getting more and more significant when the order is 7 or greater. Several variances couldn't be computed for the larger orders, as the assay was not able to pick up the tiny amounts of RNA when background RNA is present in large quantities.

are, in most cases, small. As such, only miRNA of levels that satisfies both conditions ($>$ 10,000 molecules, $<$ 7 orders lower in concentration than the background RNA) should be considered accurate.

Based on the observation that the gradients of C_t (threshold cycle) values is roughly equal across different background RNA levels (gradients can be found in the form of $y = mx + c$ of the trendlines in Figure 2.7, where m is the gradient), the effects of background RNA can be corrected by comparing the C_t of any miRNA to that of 2S rRNA, the positive control that varies in C_t with background RNA levels. When the difference of C_t values between the miRNA of interest and 2S rRNA is taken, this value should be the same irrespective of background RNA values.

The assay is specific for its target sequence — over the course of performing the RT-qPCR experiments, it was noted that no, or very little, detectable fluorescence was produced if the wrong primers were used, or if there were no RNA in the starting sample (data not shown). The latter situation is routinely carried out as a negative control, demonstrating that the

fluorescence produced in other wells is a result of the use of correct primers and the presence of the probed RNA.

2.2.4.3 Upper-limit saturation in the detection of 2S rRNA

While examining the raw fluorescence data of 2S rRNA in the manufacturer's software and LinRegPCR, it was observed that the wells probing for 2S rRNA had baselines (the "stationary phase" in the sigmoidal fluorescence curve) that were less accurate than other wells. This was due to 2S rRNA tending to cross the fluorescence threshold around 9–13 cycles due to its abundance in tissues. Common guidelines in the interpretation of qPCR data mention that analysis should only be performed on wells with C_t greater than 10. From experience, it has been observed that accuracy of C_t values under 10 suffers due to the assumptions made by PCR software, as it assumes that the fluorescence observed in cycles 2–6 is the result of noise (background reading), and accordingly reduces the fluorescence reading for all other cycles to remove this noise. However, in the case of 2S rRNA, the sheer abundance of the RNA leads to a detectable increase of fluorescence in the first few cycles, which is mistakenly treated as noise by the software. Thus, there is an upper limit to the detection of RNA concentrations — to use some values to illustrate my point, if an RNA is abundant enough to produce a C_t of 8 originally, a fourfold-increase in RNA concentrations might not even change the value of this C_t. In the absence of this saturation effect, the C_t should have been 6.

In order to check for the validity of 2S rRNA data with C_t below 10, I ran an assay to ascertain whether C_t values below 10 could be corrected to its "intended" reading. Serial dilutions of 2S rRNA across two biological replicates were run, producing a graph that confirms the saturation effect (Figure 2.9). Although equations could have been constructed to correct the readings, I decided that this correction will be introducing bias and inaccuracy into 2S rRNA readings, which is especially important it serves as a comparative baseline for all assayed miRNAs. Eventually, the entire experiment was re-run with one extra modification: wells that were probing for 2S rRNA had total RNA diluted 1,000-fold, resulting in C_t values of about 22–30 cycles.

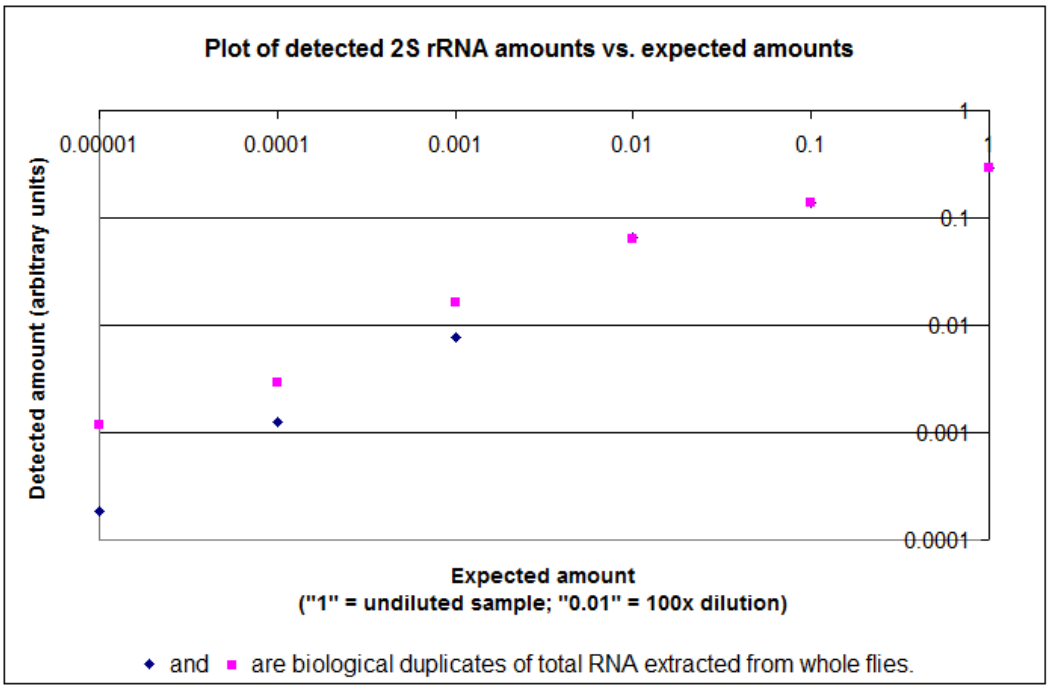


Figure 2.9: Plot of detected values versus expected values of 2S rRNA. Theoretically speaking, all six points should lie on a straight line (with a gradient that is related to the amplification efficiency). However, the three points with the highest concentrations (0.01 to 1) that corresponds to threshold cycles of 9–15 shows a small degree of saturation in detection.

2.2.4.4 Calculation of relative fold change in miRNA assays

The Pfaffl method (Pfaffl, 2001), which is a slight modification of the Livak method (Livak and Schmittgen, 2001), was used to calculate expression ratios for the assayed miRNAs in specific tissues relative to whole fly levels. A demonstration of the method with hypothetical C_t values is shown in Table 2.6.

	C_t values	
	Tissue X	**Whole fly**
miRNA-Y	16	17
2S rRNA	15	13
ΔC_t (miRNA Y - 2S rRNA)	1	4
$\Delta\Delta C_t$ (tissue X - whole fly)	-3	
Relative expression ratio (tissue X ÷ whole fly)	$2^{-(-3)} = 8$ (Livak method) $1.85^{-(-3)} = 6.3$ (Pfaffl method)	

Table 2.6: Demonstration of the calculation of relative expression ratios using the Livak and Pfaffl methods. Both methods are largely similar, except for the final step, where different PCR efficiency values are used.

For the Livak method, the underlying assumption is that the template amplification during qPCR is 100% efficient (i.e. doubling of template and fluorescence every cycle) for all reactions; for the Pfaffl method, a value between 1–2 is chosen as an approximation of the qPCR efficiency across all reactions. A value of 1 implies that the qPCR is 0% efficient, while 2 implies that the qPCR is 100% efficient.

In my experiments, I have assayed serial dilutions of the synthetic 26-mer RNA and 2S rRNA. The qPCR efficiencies in those experiments tended to range between 80–90% (data not shown). As such, I opted to use 1.85 (85% efficiency) for the calculations of relative expression ratios. This value is somewhat in agreement with existing literature — Karlen *et al.* (2007) reported that the individual efficiencies from 704 PCR reactions had an approximately normal distribution, with values ranging from 1.4 to 2.15 with a peak value around 1.85.

One of the biggest assumptions for the Pfaffl method is that reaction efficiencies are constant across the different primers used in the qPCR experiment. This assumption has been tackled head-on by more sophisticated algorithms that allow for the estimation of reaction efficiencies directly from the raw data, bypassing the need of making serial dilutions to cal-

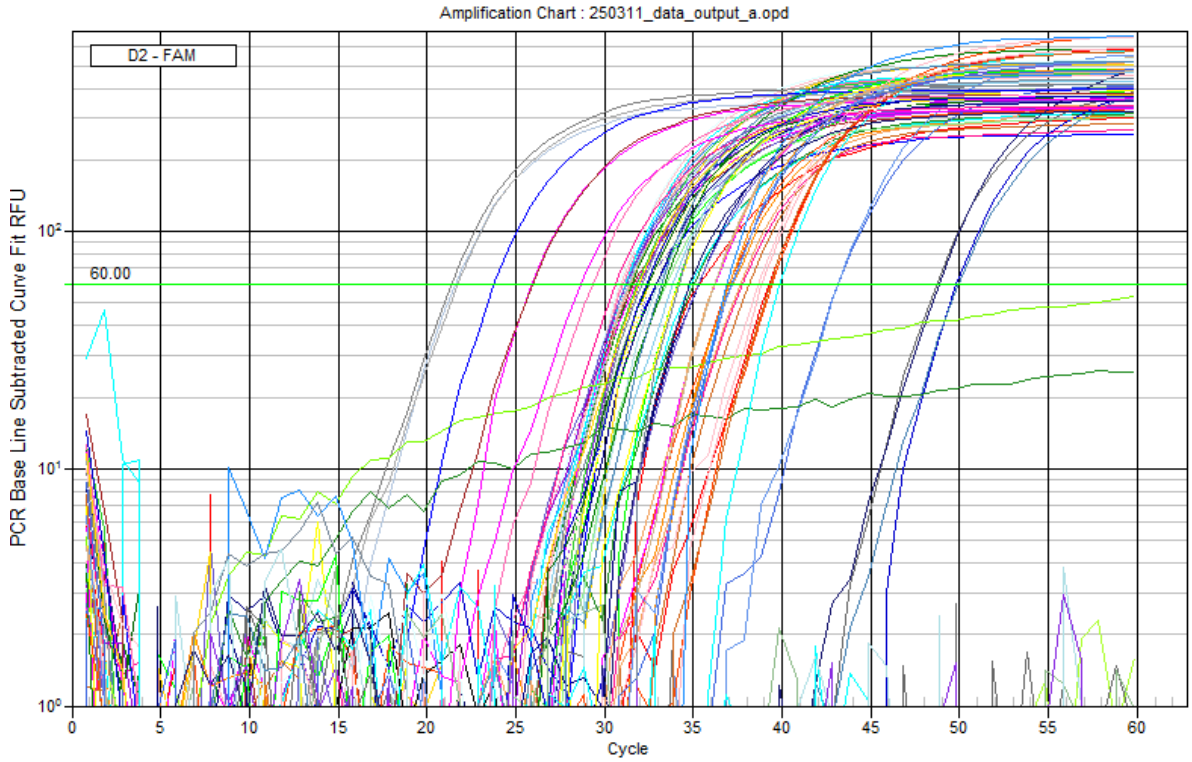


Figure 2.10: An example of a qPCR run on a 96-well plate, probing for 10 miRNAs and 2S rRNA in adult crop. Note that the y-axis is logarithmic. The rate of increase in fluorescence for individual reactions during their log phases is largely similar, supporting the assumption that reaction efficiencies can be approximated to a universal value. The two wells without a distinct log phase are negative controls.

culate reaction efficiencies for specific primers. An example of such a program is LinRegPCR (Ruijter *et al.*, 2009). However, LinRegPCR was not able to calculate reaction efficiencies automatically from my data, often requiring user input to produce results. As this input can change the resulting reaction efficiencies significantly, unwanted bias is introduced into the reaction efficiency calculations from my qPCR results.

I felt that a universal PCR efficiency value is a reasonable assumption for my data for several reasons. My assays probe for templates of very similar length (21–23 nt for miRNAs, 26 nt for the synthetic RNA, and 30 nt for 2S rRNA), hence the time required for extension during qPCR would be similar for all templates. Also, raw data from the qPCR experiments show that the individual reactions have similar gradients in its log phase (the rate of product accumulation when the fluorophore is not limiting), implying that PCR efficiencies are similar across reactions. An example of this observation is illustrated in Figure 2.10.

2.3 Results and Discussion

2.3.1 Significant tissue-miRNA pairs

In a given tissue, the genes that are targeted by a miRNA can either be overexpressed, underexpressed or not significantly different from whole fly levels. When considering a list of overexpressed or underexpressed genes in a tissue, the proportion of these genes being targeted by the miRNA in question can be either unexpectedly large or unexpectedly small. An analogy can be drawn using egg yolks: a list of overexpressed genes in tissue X (the egg) will have a fraction of genes targeted by miR-Y (the yolk). The yolks can sometimes be larger than expected (unexpectedly large proportion of genes targeted by miR-Y) or smaller (unexpectedly small proportion of genes targeted by miR-Y). The probability of the "yolk" being a certain size is calculated using Fisher's exact test, and if the P value is less than 0.05 (post-Benjamini-Hochberg correction), the miRNA affecting the tissue-specific genes is referred to below as a "significant tissue-miRNA couple".

My early computational efforts (Liew and Micklem, 2008) (unpublished) identified 65 significant tissue-miRNA couples — 11 for underexpressed genes and 54 for overexpressed ones (the full list of predictions are in the Appendix, Table TODO). For both the underexpressed and overexpressed genes, there were 6 tissue-miRNA couples that occurred in both lists. These 6 couples are known as "tissue-miRNA pairs". Gorton and Micklem (2009) (unpublished) performed calculations on an expanded set of FlyAtlas data that included 5 extra tissues with a more permissive cutoff, and it produced a similar list of 17 tissue-miRNA pairs. Their predictions and mine are compared in Table 2.7.

The expected scenario for these tissue-miRNA pairs is that in terms of the list of overexpressed genes, there should be a significant depletion of genes targeted by the miRNA in question; in terms of the list of underexpressed genes, there should be a significant enrichment of genes targeted by the miRNA. This is because the expected mechanism of miRNAs is to downregulate gene expression in a post-transcriptional manner. To our surprise, the observed situation for all 17 pairings is the exact opposite of this: there is a significant enrichment

miRNA	Liew **and** Micklem **(2008)**	Liew **and** Micklem **(2008), corrected**	Gorton **and** Micklem **(2009)**
let-7	ovary	ovary	ovary
miR-1			ovary
miR-1013	ovary	ovary	ovary
miR-11	adult carcass head hindgut	adult carcass hindgut	adult carcass head hindgut larval hindgut virgin spermatheca crop
miR-2a	head		
miR-277			ovary larval hindgut larval salivary gland
miR-79	ovary		ovary
miR-92a	head		
miR-92b	head adult carcass	head adult carcass	head adult carcass virgin spermatheca
miR-iab-4-3p	ovary		ovary

Table 2.7: Overview of my initial predictions against Gorton and Micklem (2009). The ten assayed miRNAs are predicted to be depleted in the tissues listed above. Six of the predictions are common across all three sets of predictions.

of genes targeted by the miRNA in the list of overexpressed genes, and vice versa for the underexpressed genes.

To ensure that this observation was not just due to chance, the cutoff for significance was raised so that more pairings were observed to occur in both lists. With a less stringent cutoff, only 12 tissue-miRNA pairs were found to follow the expected scenario, while 65 others were found to follow the "counterintuitive" scenario noted above.

The most plausible explanation for this observation is that miRNAs show tissue specificity by being downregulated in that tissue. In this light, the "counterintuitive" scenario makes sense — if the miRNA is predicted to be repressed in the tissue in question, overexpressed genes would consist of a unexpected enrichment of mRNA targeted by the miRNA, and vice versa for the underexpressed genes. Experiments were designed to investigate this hypothesis.

Another less likely, but plausible, hypothesis is that the miRNAs might be activating gene expression of their respective mRNA targets. This phenomenon, termed RNAa (RNA activation), has been expounded in detail in the introduction. In this case, the miRNA is predicted to be overexpressed in the tissue.

2.3.2 Preliminary miRNA assays confirming hypothesis

After performing test miRNA assays on whole Oregon R flies, assays were carried out using Canton S flies (from the FlyAtlas lab). It is predicted that for every tissue-miRNA pair identified through computational methods described above, the miRNA is depleted in that tissue, relative to whole fly levels.

Preliminary experiments were performed on tissue-miRNA pairs that involved either the head or ovary, due to their relative ease of dissection and as they accounted for about half of the predictions. If the miRNA was predicted to be underexpressed in the ovary, it should be roughly at whole fly levels in the head, and vice versa. Table 2.8 summarises the assay results of 10 miRNAs in two tissues.

As seen in Table 2.8, it can be seen that most (seven out of nine) of the probed miRNAs confirm the hypothesis. There are some cases (miR-92a, miR-92b) where the miRNA is

miRNA	Ratio of expression			Predicted to be depleted in...	Fits predictions?
	Head	Ovary	Whole fly		
let-7	2.523	0.061	1.000	ovary	Yes
miR-1	1.238	0.007	1.000	ovary	Yes
miR-1013	0.020	0.069	1.000	ovary	Yes
miR-11	2.783	0.367	1.000	adult carcass head hindgut larval hindgut virgin spermatheca crop	 No
miR-2a	3.184	0.571	1.000	head	No
miR-277	7.314	0.002	1.000	ovary larval hindgut larval salivary gland	Yes
miR-79	1.752	0.385	1.000	ovary	Yes
miR-92a	0.072	0.885	1.000	head	Yes
miR-92b	0.009	0.470	1.000	head adult carcass virgin spermatheca	Yes
miR-iab-4-3p	N/A	N/A	N/A	ovary	N/A

Legend

: tested

Table 2.8: Overview of miRNA assays on dissected tissues. Ratios of expressions are calculated from the average of two biological replicates. miR-iab-4-3p is reported to be present in the embryo but not in adult tissue in a microarray study (Ruby *et al.*, 2007), and this is confirmed in these experiments.

depleted in both tissues. For these miRNAs, it is possible that they are upregulated in other tissues not dissected in this preliminary study, thus producing a higher whole fly average. This early success led to the assays of these miRNAs in more tissues to further test the hypothesis.

2.3.3 Assays of 10 miRNA across 12 fly tissues

Following the early success, RT-qPCRs were carried out to assay the levels of the 10 miRNAs of interest across 12 fly tissues. The dissection of these tissues is described in more detail in

Section 2.2.3. Eight of those tissues were from adult flies: head, ovary, testis, tubule, midgut, hindgut, crop and male accessory gland; while the remaining four were from fly larva: fat body, tubule, midgut and hindgut.

2.3.3.1 miRNAs with expression patterns that fit predictions

By defining significant depletion of the miRNA in the dissected tissues as two-fold underexpression relative to whole fly levels, the bar chart in Figure 2.11 shows six predicted depletions (marked with red asterisks) that were borne out by experimental results. With the exception of the depletion of miR-92b in adult heads, the upper bound of the 95% confidence intervals of the other five were at least two-fold depleted relative to whole fly levels.

The accuracy of this assay can be assessed by comparing the results to published literature. For let-7, its temporal expression in *Drosophila* has been well-studied due to its role in remodeling the abdominal musculature during the larval-to-adult transition (Sokol *et al.*, 2008). Mutations in let-7 lead to severe reduction in fitness — Sokol *et al.* (2008) reports that ∼43% mutants died prematurely during the course of development, the majority of which arrested at the end of metamorphosis; Caygill and Johnston (2008) found that a majority of their let-7 mutants were unable to eclose, and those that manage to eclose had severely shortened lifespans. During larval development, the expression of let-7 is triggered by a pulse of ecdysone prior to pupal formation (Sempere *et al.*, 2002).

The expression pattern of let-7 during the larva-to-adult transition, as well as in adult ovaries, is shown in Figure 2.12. As expected, the levels of let-7 in the larval tissues and adult ovaries in my assay were significantly depleted relative to adult whole fly levels. Interestingly, in my assay, I quantified the expression of let-7 to be 0.21 that of whole flies, which is very similar to the value of 0.25 obtained by Sempere *et al.* (2002). It is likely that the depletion of let-7 in ovaries resulted in the tissue-specific pattern of expression of its targets, i.e. higher-than-average number of let-7 targets being upregulated, and lower-than-average number of let-7 targets being downregulated.

Like let-7, miR-1 is another miRNA that is conserved across most metazoans. The tissue-

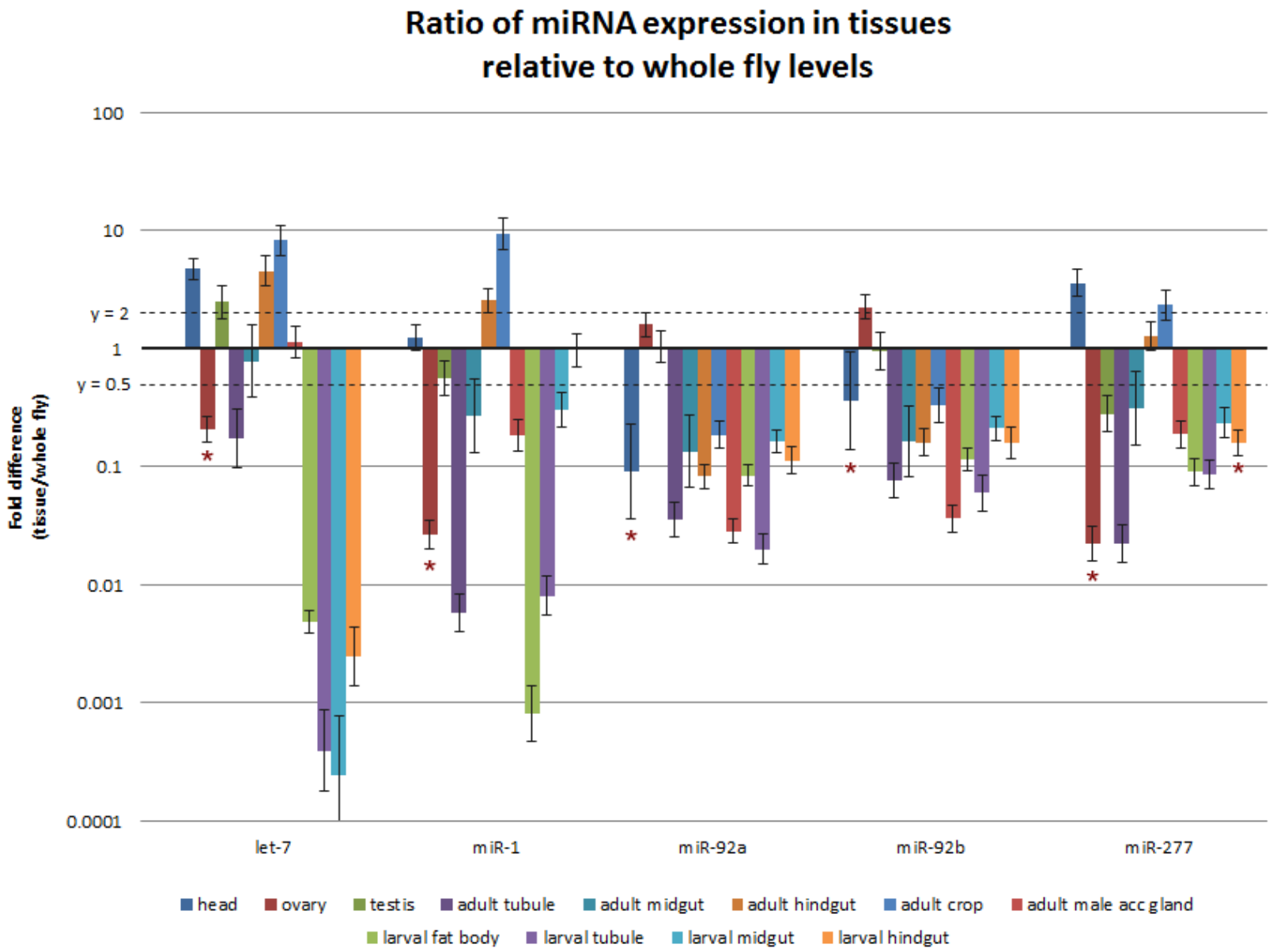


Figure 2.11: Bar chart showing ratio of expression of let-7, miR-1, miR-92a, miR-92b, and miR-277 across 12 fly tissues. Red asterisks indicate the predicted depletion of the miRNA in that tissue. Error bars indicate 95% confidence interval for the values. Dotted lines indicate two-fold increase or decrease in expressions.

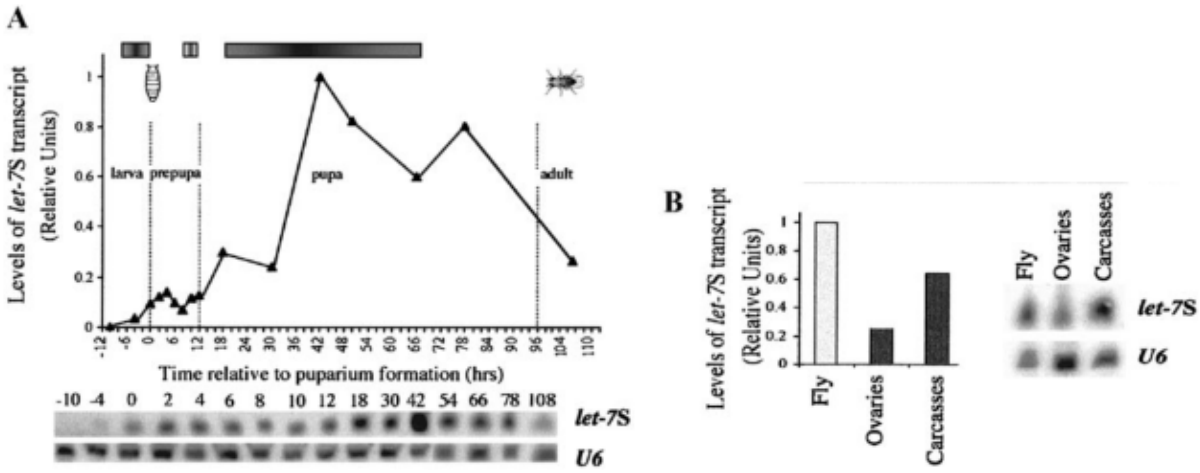


Figure 2.12: Northern blot analysis of let-7 expression during the development of *D. melanogaster* (A, left), and in adult ovaries (B, right). As let-7 promotes the transition from the larval to adult stage, its expression is first detected at the late third instar larval stage, and reaches a maximum in the second day of the pupal stage. In adults, the expression of let-7 in the ovaries is estimated to be 0.25 of whole fly levels. Both figures are from Sempere *et al.* (2002).

specific pattern of expression is phylogenetically conserved as well. As in zebrafish, mice and humans, *Drosophila* miR-1 is highly expressed in the mesoderm of early embryos. Interestingly, the growth of *Drosophila* miR-1 mutants arrests at the transition from the first to second instar larval stage, demonstrating that miR-1 is not required for embryonic development, but it is required for the post-mitotic growth of larval muscle (Sokol and Ambros, 2005). Temporally, the expression of miR-1 has been shown to gradually increase during embryonic development, and the expression persists into larval and adult stages (see Figure 2.13).

From my data, the observed depletion of miR-1 in adult ovaries, relative to whole fly levels, is likely due to the exclusion of maternal miR-1 from the developing oocytes. As with let-7, the targets of miR-1 had a tissue-specific expression pattern that was predicted as significant by my computational methods.

While the biological function of miR-92a and -92b in *Drosophila* remain unclear, both miRNAs could be co-transcribed due to their proximity (5 kb apart) on the same strand of DNA. The transcription of miR-92a is driven by a protein coding gene *jigr*, as miR-92a is located in one of the introns of the gene. In a large scale small RNA sequencing performed by

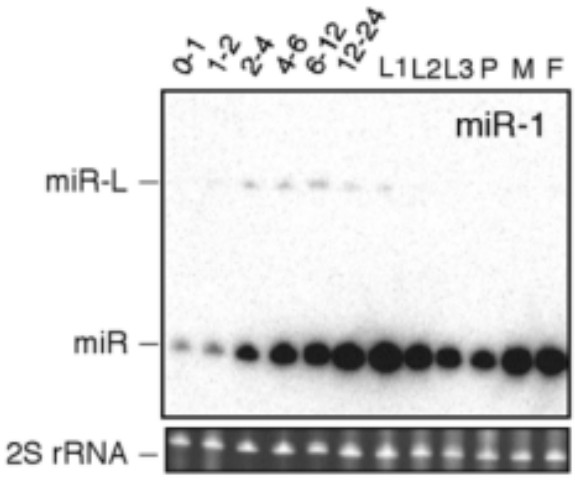


Figure 2.13: Northern blot of miR-1 for *Drosophila* at different stages of development. Development stages of embryos are noted as hours (0-24) after egg laying. L1, L2 and L3 denote the larval stages; P the pupal stage; while M and F as adult males and females respectively. Figure adapted from Aravin *et al.* (2003).

Ruby *et al.* (2007), the expression profiles of miR-92a and -92b across ten different datasets were very similar to each other. This observation holds true in my assays as well — miR-92a and -92b show very similar expression patterns across all dissected tissues. The significance underlying the depletion of miR-92a in the head remains unclear.

miR-277 has been shown to be expressed at higher quantities in the adult stage. Two studies did not detect any miR-277 expression during embryonic nor larval stages (Aravin *et al.*, 2003; Lai *et al.*, 2003), while another detected low levels of miR-277 expression in both stages (Ruby *et al.*, 2007). As many of the predicted targets of miR-277 are involved in the catabolism of valine, leucine and isoleucine, Stark *et al.* (2003) predicted that miR-277 acts as a metabolic switch that regulates metabolic responses to environmental stresses. In my assays, the expression of miR-277 is consistently low across all four dissected larval tissues, including the larval hindgut, which has been predicted to have low miR-277 amounts from the tissue-specific expression patterns of its targets. The functional significance of the depletion of miR-277 in adult ovaries is not known, but there is at least one study that concurs with the expression pattern of miR-277 in fly heads and ovaries. Berezikov *et al.* (2011) reports that miR-277 represents 0.08% of all miRNAs in the ovaries; this figure jumps to 4.24% in fly heads.

miRNA	0–1 h embryos	2–6 h embryos	6–10 h embryos	12–24 h embryos	Larva	Imaginal discs	Pupae	Adult heads	Adult bodies	S2 cells
miR-2a	9.56	8.03	16.79	15.00	7.07	10.21	7.32	7.82	5.41	12.79
miR-11	6.58	4.73	7.32	9.17	12.68	8.49	12.57	3.61	8.93	25.92
miR-79	24.59	8.98	7.28	3.08	9.77	21.16	8.52	3.16	4.12	9.35

Table 2.9: Normalised expressions of miR-2a, miR-11 and miR-79 across ten datasets. There was no data on miR-1013. Across the datasets, the largest values are within tenfold of the smallest one. Data obtained from Ruby *et al.* (2007).

2.3.3.2 miRNAs with expression patterns that did not fit predictions

As shown in Fig 2.14, the ratios calculated from the RT-qPCR assays do not support the predicted depletions of miR-2a, miR-11, miR-79 and miR-1013 in the tissues marked with red asterisks. The extent of depletion of miR-1013 in adult ovaries is insufficient to be considered significant. Compared to the expression patterns in the previous bar graph (Figure 2.11), the expression of these four miRNAs is more even across all tissues. All, bar one, of the calculated ratios are within ten-fold of the whole fly levels. This observation is somewhat supported in literature — Table 2.9 shows the normalised small RNA reads across multiple developmental time points, as summarised from Ruby *et al.* (2007).

From my assays, the overexpression of miR-2a in adult heads relative to ovaries is at odds with the RNAseq data from Berezikov *et al.* (2011). In their experiments, miR-2a was 11.4% of the total ovarian miRNAs, while this value is at a much lower 0.58% of all head miRNAs — these values correspond to my initial computational predictions much better than my assay results. Nonetheless, judging from raw values, miR-2a was almost always the most abundant miRNA among the ten assayed across my tissues, which fits the high abundances of miR-2a observed in Berezikov *et al.* (2011).

It is interesting to note that for miR-11, all of the predicted depletions turned out to be enrichments in the specific tissues. It might be that miR-11 is directly activating the expression of its targets (RNAa) instead of repressing it. However, there is no evidence in the literature to support such a claim, as the biological role of miR-11 in *Drosophila* remains unclear.

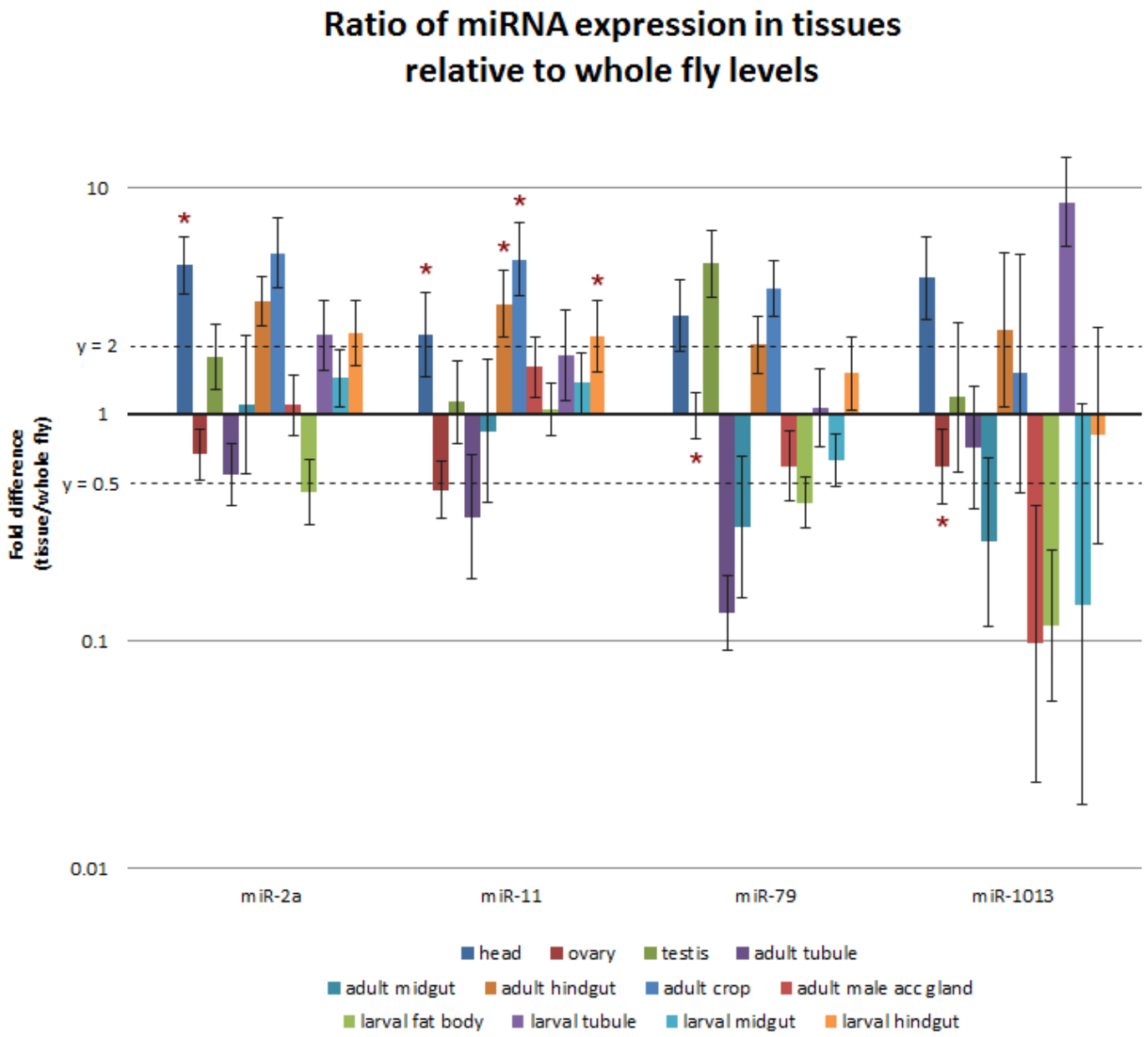


Figure 2.14: Bar chart showing ratio of expression of miR-2a, miR-11, miR-79 and miR-1013 across 12 fly tissues. Red asterisks indicate the predicted depletion of the miRNA in that tissue. Error bars indicate 95% confidence interval for the values. Dotted lines indicate two-fold increase or decrease in expressions.

2.3.3.3 miR-iab-4-3p

The results for miR-iab-4-3p were inconclusive due to the extremely low amounts of the miRNA in all dissected tissues. Approximately half of the wells did not produce any detectable fluorescence, while nearly all of the remaining reads had C_t values of > 45. This observation is supported by Ruby *et al.* (2007), as miR-iab-4-3p was completely absent in the larval and adult RNAseq datasets.

As seen in Fig 2.15, the highly variable readouts resulted in a very large 95% confidence interval for the ratio of expressions. Although the data appears to be consistent with the predicted depletion in ovaries, further experiments are needed to confirm the veracity of this observation.

2.4 Conclusions

Despite uncovering the identities of hundreds of miRNAs in *Drosophila*, the biological role for many remain unclear. In this chapter, the tissue-specific expression of miRNAs was predicted from the tissue-specific expression of its targets: 10 miRNAs were predicted to be depleted across several fly tissues, with the hypothesis that miRNAs show tissue-specific activity by being depleted in said tissue.

To verify these predictions, RT-qPCR assays were run on 12 fly tissues, 8 of which were from adults and 4 from larva. Prior to establishing a collaboration to dissect all these tissues, I ran experiments to show that the RT-qPCR protocol used by Varkonyi-Gasic *et al.* (2007) to assay plant miRNAs work well on fly tissues as well. The limits of accuracy, sensitivity and specificity of the assay was tested to ensure that subsequent results were reliable and reproducible.

From the results of the assay, about half of the predicted depletions were *bona fide* — the depletion of let-7 and miR-1 from adult fly ovaries can be attributed to the observed absence of both miRNAs in oocytes, but the biological significance that underlie the tissue-specific depletions of miR-92a, miR-92b and miR-277 is unclear, as the roles of these miRNAs in flies

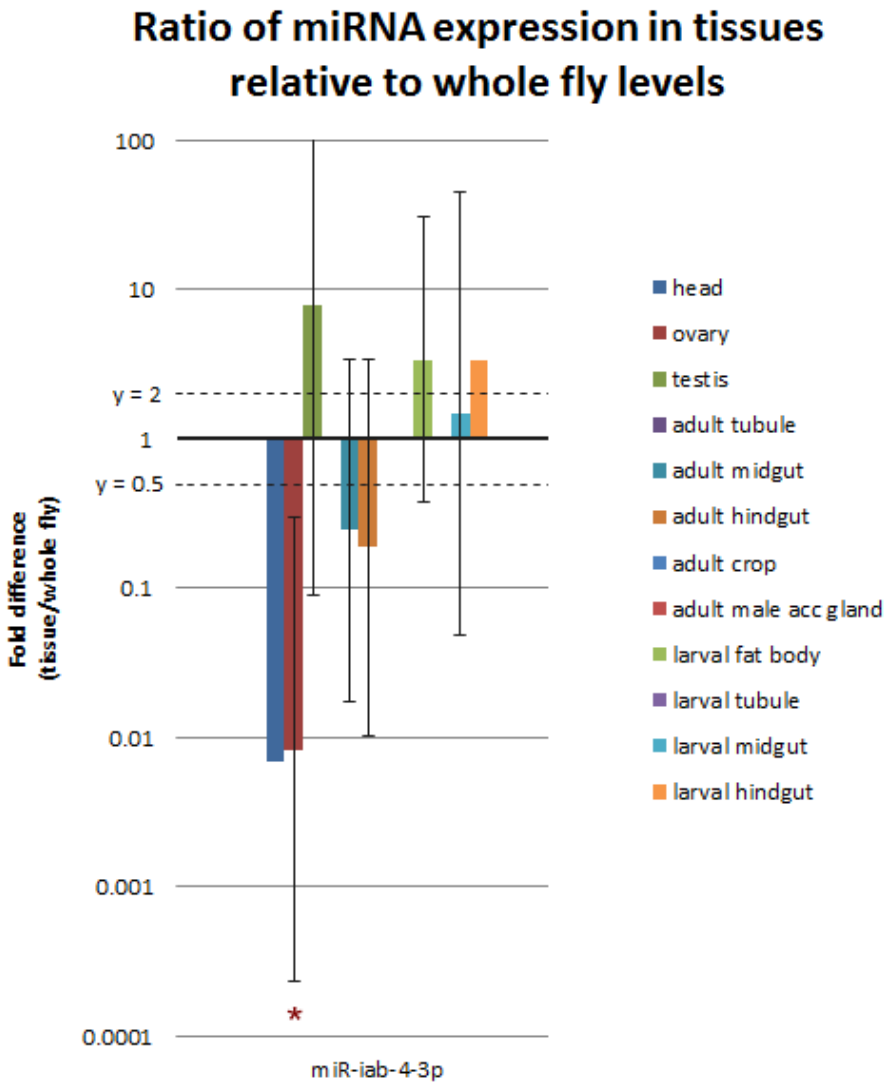


Figure 2.15: Bar chart showing ratio of expression of miR-iab-4-3p across 12 fly tissues. Red asterisks indicate the predicted depletion of the miRNA in that tissue. Error bars indicate 95% confidence interval for the values, which can only be calculated if at least two out of three technical replicates results in detectable fluorescence.

have yet to be experimentally verified.

For miR-2a and miR-11, the tissue-specific predicted depletions turned out to be enrichments in those tissues instead. As the predicted depletions of miR-11 in all four tissues turned out to be enrichments, it is possible that miR-11 activates the expression of its targets, instead of repressing them. For miR-79 and miR-1013, the expression levels of both miRNAs in the ovary were not significantly different from whole fly levels.

Lastly, due to the extremely low amounts of miR-iab-4-3p in the dissected tissues, the results that were produced from the assay were unreliable. A more sensitive assay is needed to verify the tissue-specific expression of miR-iab-4-3p.

2.5 Future work

As shown in our results, the method used to predict tissue-specific depletions of miRNA has achieved limited success. Additional constraints, such as filtering out mRNA targets that are regulated by multiple miRNAs, or only using data from mRNAs that are expressed above a certain level in tissues, will allow the algorithm to focus on miRNA-mRNA targeting that is unique to the miRNA with a more discernible regulatory effect. In the future, the predicted miRNA-mRNA targeting dataset (MicroCosm Targets) can be directly substituted with data from experimentally verified mRNA targets of fly miRNAs, which will result in biologically meaningful tissue-specific expression patterns for some miRNAs.

It would be interesting to find out whether the tissue-specific depletions of miRNAs is related to the up- or downregulation of specific pathways in said tissue. However, pathway analysis checking for the enrichment of miRNA targets in KEGG Pathways (Ogata *et al.*, 1999) did not produce any meaningful results. Here, the use of experimentally verified mRNA targets of miRNAs instead of a predicted database might be key in getting biologically meaningful results from the pathway analysis.

With sufficient funding and manpower, an ambitious project that can be undertaken would be to obtain small RNAseq data from all the dissected tissues in FlyAtlas. That way, direct correlations can be made between the expression levels of miRNAs and their mRNA

targets. The tissue-specific expression patterns observed in the small RNAseq data would also be invaluable to researchers that are trying to understand the biological roles of their miRNAs of interest. The main barrier to this idea — assuming funding is not an issue — is the immense numbers of smaller tissues that would have to be dissected for library construction. However, as single-molecule sequencing technologies are starting to mature and will be more affordable in the future, both the amount of RNA required for sequencing and the cost to do so will be feasible in a few years' time.

Chapter 3

Optimising RNA extraction from *Symbiodinium sp.* cultures for RNA-Seq

3.1 Introduction

RNA-Seq is a method that profiles transcriptomes using next-generation sequencing platforms. Despite only being developed roughly five years ago, many researchers are favouring RNA-Seq over existing hybridisation-based methods to quantify changes in transcript expression levels, as the former method is more accurate and senstive than the latter (Wang *et al.*, 2009).

Broadly speaking, an RNA-Seq experiment can be divided into two parts: the production of a cDNA library from a population of RNA by reverse transcription, followed by the sequencing of the library using the next-generation platform of choice. The reverse-transcribed RNA is not just limited to total RNA — selective enrichment of a subset of RNA can be carried out depending on the aim of the experiment. For instance, poly(A)+ selection can be carried out to investigate changes in mRNA expression levels, or perhaps the small RNA fraction can be enriched to study miRNAs. The resulting library can then be sequenced via high-throughput methods available to the researcher. Figure 3.1 illustrates a typical RNA-Seq experiment that focuses on changes in mRNA expression levels.

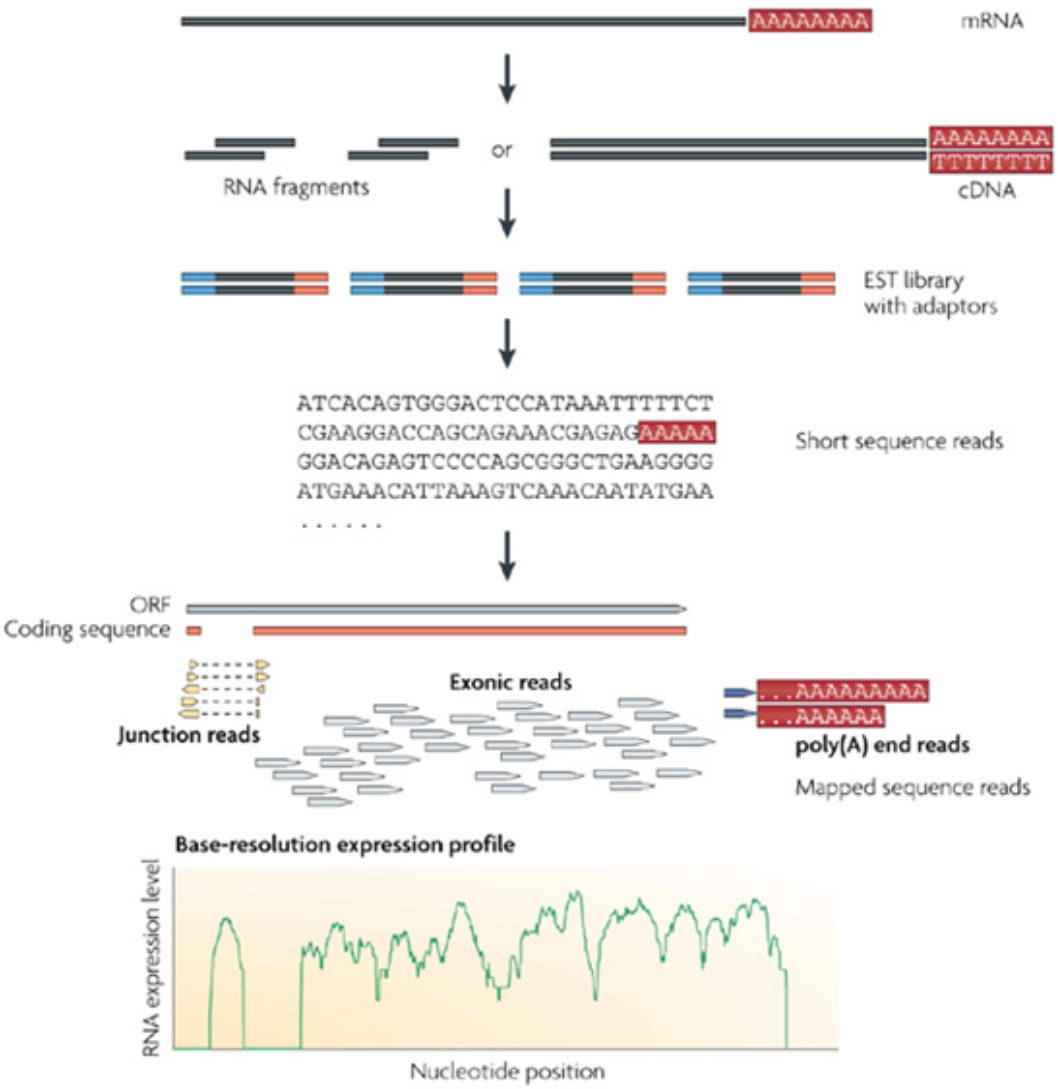


Figure 3.1: Overview of a RNA-Seq experiment. Libraries are generated by adding sequencing adaptors (blue and red) to cDNA fragments. Read data produced by the sequencing run can then be mapped back to the reference genome, allowing the expression of the transcript to be quantified in a digital manner. Illustration is adapted from Wang *et al.* (2009).

3.1.1 Comparison between RNA-Seq and existing hybridisation-based methods

As RNA-Seq does not require the prior knowledge of the genomic sequence, it allows for the detection of a wider range of transcripts than hybridisation-based methods. This advantage is especially useful in profiling transcriptional changes in non-model organisms, as the genome might not be available, or incomplete, for the organism of interest (Wang *et al.*, 2009).

Also, unlike hybridisation methods, the upper or lower limit to the detection of transcripts is far less affected by saturation effects (at high expression levels) or by noise (at low expression levels). For example, the dynamic range where expression levels are accurately quantified in DNA microarrays is much smaller than that of RNA-Seq: a few hundredfold for the former, while almost 10,000-fold for the latter (Wang *et al.*, 2009). RNA-Seq data has been shown to be in better agreement with data from qPCR, and also being more reproducible across biological and technical replicates (Nagalakshmi *et al.*, 2008).

Given sufficient sequencing depth, RNA-Seq is able to discover SNPs in the sequenced transcripts at higher sensitivities, unlike in DNA microarrays where highly related sequences are prone to cross-hybridise (Shendure, 2008). For example, Cloonan *et al.* (2008) predicted 2,000 SNPs in mice embryonic stem cells and embryoid bodies from approximately 10 gigabases of transcriptomic data. Roughly a third of these predictions matched SNPs previously reported in RefSeq (NCBI **Ref**erence **Seq**uences, `http://www.ncbi.nlm.nih.gov/RefSeq/`) and dbSNP (NCBI Single Nucleotide Polymorphism database, `http://www.ncbi.nlm.nih.gov/snp`).

3.1.2 Methods to assess quality of extracted RNA

Obtaining high quality RNA is important for producing reliable results from sequencing methods. Traditionally, the quality of extracted RNA samples is assessed by running it on an agarose gel, and comparing the intensities of the resulting 18S and 28S rRNA bands. If the 28S/18S ratio exceeds 1.8, the sample is deemed to be high quality (Sambrook and Russell, 2001; Copois *et al.*, 2007). However, as this method relies on the subjective

interpretations of gel images, results might differ from one lab to another (Schroeder *et al.*, 2006).

To avoid ambiguity, we sought to determine the quality of our extracted RNA using a Bioanalyzer 2100 machine, which calculates a RIN (RNA integrity number) value based on the electropherogram produced by the samples. RIN values range from 1 to 10: 1 being completely degraded RNA, while 10 being perfectly intact RNA (Schroeder *et al.*, 2006). RIN values have been shown to be better than 28S/18S ratios at discriminating samples that ended up producing reliable microarray data (Copois *et al.*, 2007). While there are varying "rule-of-thumbs" regarding RIN value thresholds to determine whether an extracted RNA sample has the sufficient quality for downstream sequencing, based on Copois *et al.* (2007)'s paper, we opted to use 7.8 as the threshold for sufficiently good quality RNA, as opposed to the lower threshold of 5–6 used by Rosic and Hoegh-Guldberg (2010). Our RIN value threshold of 7.8 is fairly close to the manufacturer's (Illumina) recommended threshold of 8 for library generation.

Additionally, we also measured the 260/230 and 260/280 ratios of the RNA samples with a NanoDrop 2000c machine as an indication of the purity of the extracted RNA samples. It is commonly accepted that a sample can be considered pure when both ratios have values of above 1.8. However, unlike RIN values, we did not strictly exclude RNA samples that fell short of either ratio thresholds, as spectophotographic measurements obtained from NanoDrop tend to be much more variable than RIN values obtained from the Bioanalyzer machine.

3.1.3 Challenges to RNA extraction unique to *Symbiodinium sp.* cells

Unlike typical eukaryotic cells, free-living *Symbiodinium* cells have a thick cell wall composed of multiple layers, as observed under an electron microscope. While the exact chemical composition of the *Symbiodinium* cell wall has yet to be determined, similarities to other dinoflagellates have led some to postulate the presence of sporopollenin in the *Symbiodinium*

cell wall (Wakefield *et al.*, 2000). Also, it has been shown that isolated cell walls of *Symbiodinium sp.* are prone to cellulose digestion, while intact *Symbiodinium* were resistant to cellulase action, indicating that cellulose might exist in the inner layer of *Symbiodinium* cell walls (Markell *et al.*, 1992).

Unfortunately, the literature on methods to overcome the thick cell wall in extracting high quality RNA from *Symbiodinium sp.* has been sparse. Mechanical efforts in disrupting the cell wall of *Symbiodinium* generally falls into two categories: grinding the sample in a mortar and pestle (Santiago-Vázquez *et al.*, 2006; Boldt *et al.*, 2009; Rosic and Hoegh-Guldberg, 2010), or high-speed agitation of the cells mixed with glass beads (Rosic and Hoegh-Guldberg, 2010). Our initial optimisation efforts centered on the latter technique, as Rosic and Hoegh-Guldberg (2010) were able to efficiently extract high quality RNA using bead-beating techniques. However, after experimenting with bead-beating methods, we discovered that grinding generally produced better quality RNA than mechanical agitation with glass beads. This is described in greater detail in Section 3.2.3.

3.2 Materials and methods

3.2.1 Culture conditions of *Symbiodinium sp.* samples

Free-living *Symbiodinium sp.* cultures of CCMP Bigelow strain CCMP2467 were used, which were originally isolated from its *Stylophora pistillata* host at Aqaba, Jordan. The dinoflagellates were cultured at 23 °C in f/2 medium (Guillard and Ryther, 1962) on a 12:12h light regime (daytime: 6 am to 6 pm; night-time: 6 pm to 6 am, light intensity of 80 µmol m^{-2} s^{-1}). The salt content in the medium is set to 40 g/L, which matches the above-average salinity characteristic of the Red Sea.

By default, exponentially-growing cells were harvested approximately at noon, which is at the middle of the cultures' daytime phase. As we were interested in investigating the transcriptional changes in *Symbiodinium sp.* when subjected to different environmental stresses, eight other temporary growth conditions were devised to mimic these stresses. In

all cases, separate exponentially-growing *Symbiodinium sp.* cultures were subjected to the conditions listed in Table 3.1, and harvested at the end of those conditions.

Stress	Details	Labelled as
Extreme cold stress	4 °C for four hours	4C
Cold stress	16 °C for four hours	16C
Heat stress	36 °C for four hours	36C
Prolonged heat stress	34 °C for twelve hours	HS
Hyposaline stress	20 g/L NaCl medium for four hours	20g
Hypersaline stress	60 g/L NaCl medium for four hours	60g
Dark stress	Substituted twelve-hour light period with darkness	DS
Dark cycle	Cultures harvested at midnight	DC
No stress	Cultures harvested at noon	noon

Table 3.1: List of the nine different conditions for our *Symbiodinium sp.* cultures. The labels denoting these conditions are used more extensively in the next chapter (the analysis of *Symbiodinium sp.* small RNAome).

3.2.2 Growth rates of *Symbiodinium sp.* samples

Samples of the growing *Symbiodinium sp.* culture were taken every few days in order to estimate the growth rates of the culture at different phases. Cell densities were measured using a haemocytometer, and the resulting values were plotted on a graph (see Figure 3.2).

The "estimated cell count" curve was fitted using the following Gompertz function:

$$\text{Estimated cell count} = 1,400,000\, e^{-15e^{-0.25t}}$$

where t is the time (in days) since the culture started growing.

Based on the graph in Figure 3.2, growth rates were calculated for the early log phase, late log phase and stationary phase (shown in Table 3.2). These values were in agreement with those reported previously (Domotor and D'Elia, 1984).

3.2.3 RNA extraction using bead-beating methods

Our initial protocol was based on earlier work done by Rosic and Hoegh-Guldberg (2010), in which they obtained high-quality RNA by homogenising their cells in a bead beater (Mag-Nalyser, Roche Diagnostics) with glass beads (0.7–1.2mm G1152 Glass beads, acid-washed,

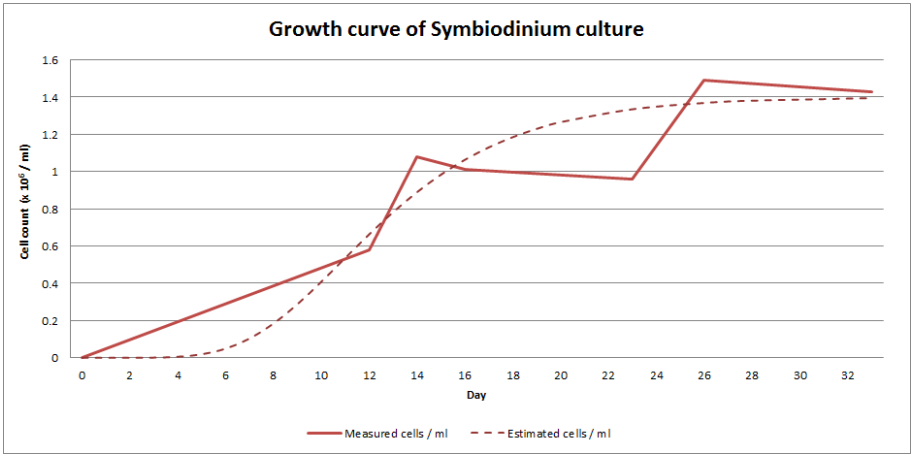


Figure 3.2: Growth curve of *Symbiodinium sp.* culture. The "estimated cell count" curve was fitted using the following Gompertz function: Estimated cell count = $1,400,000\, e^{-15e^{-0.25t}}$, where t is the time (in days) since the culture started growing.

Phase	Day	Mean growth rate	Reported rate
Early log	Day 8–12	0.32	0.35
Late log	Day 13–17	0.09	0.19
Stationary phase	Day 18–33	0.01	< 0.05

Table 3.2: Summary of *Symbiodinium* growth rates. "Reported rate" refers to growth rates published in Domotor and D'Elia (1984).

Sigma). The greatest RNA yield was produced by shaking cells at 4,500 rpm for 180 seconds. After homogenisation, the best quality RNA was achieved by TRIzol extraction (Invitrogen), followed up by a subsequent clean-up step using the RNeasy Mini Kit (Qiagen).

Using their protocol as a starting point, we were interested in creating an optimised protocol that worked best on our *Symbiodinium sp.* cultures. As efficient disruption of the *Symbiodinium sp.* cells is crucial in obtaining high-quantity and high-quality RNA, three variables were tested in the homogenisation step – shaking speed, shaking time and bead size. Bead size, in particular, seems to be an important variable that has not been tested before. Biospec (`http://www.biospec.com/instructions/beadbeater/`) recommends the use of 0.1 mm beads for bacteria; 0.5 mm beads for yeast; 1.0 or 2.5 mm for chopped-up plant or animal tissue. As we were not dealing with chopped-up tissues, we decided to focus our testing on bead sizes of 0.1 and 0.5 mm.

To test these variables, equal amounts (estimated to be 5.8 x 10^6 cells) of stationary phase *Symbiodinium sp.* cultures were homogenised in bead beaters set to different speeds

(1200–4600 rpm using three different machines) and time (90–270 seconds), with the use of different sized beads (0.1/0.5 mm). RNA from the homogenised samples was extracted using QIAzol (Qiagen). The extracted RNA was then cleaned up with the RNeasy Mini Kit (Qiagen). The full protocol can be found in the Appendix (see Section 5.2.1).

After the clean-up step, extracted RNAs were quantified using a spectrophotometer (NanoDrop 2000c, Thermo Scientific). Additionally, qualities of the extracted RNA were assessed using a RNA 6000 Nano Chip in a Bioanalyzer 2100 machine (Agilent).

3.3 Results and Discussion

3.3.1 RNA extraction using bead-beating methods

The results from these extractions are summarised in Table 3.3. Most of our extracted RNA had quantities (∼1–5 µg) similar to those obtained in Rosic and Hoegh-Guldberg (2010). However, despite using approximately the same number of starting cells (10^6–10^7 cells), we were unable to replicate the steep increase of RNA yields using their optimal settings — shaking speed of 4,500 rpm and shaking time of 180 seconds — which produced 17–22 µg of RNA for them. In our experiments, quantity-wise, the highest amount of RNA (5 µg) was produced using a shaking speed of 4,600 rpm for 300 seconds; quality-wise, the highest RIN value (6.2) was achieved using a shaking speed of 2,500 rpm for 300 seconds.

Although we were able to extract RNA of qualities (RIN values ∼6.0) comparable to that obtained by Rosic and Hoegh-Guldberg (2010), the RIN values from our extractions were not above the threshold of 7.8 that we wanted to ensure reliable results from downstream applications Copois *et al.* (2007). One of the major reasons affecting the quality of our RNA was the presence of unknown degradation products, which manifests as secondary peaks between the marker peak and 18S peak on the electropherogram produced by the Bioanalyzer 2100 machine. When viewed as a gel, these extra peaks correspond to extra bands in between the marker and 18S bands (see Figure 3.3).

Initially, it was thought that the degradation was caused by the vigorous agitation of the

Machine used	Bead size / mm	Shaking speed / rpm	Shaking time / s	Sample A				Sample B			
				Quantity / µg	260/280 ratio	260/230 ratio	RIN value	Quantity / µg	260/280 ratio	260/230 ratio	RIN value
Mini BeadBeater-1 (BioSpec)	0.1	4200	180	1.35	2.08	0.16	1.3	1.36	0.14	0.08	4.4
	0.1	4600	180	1.58	2.26	0.22	2.6	1.42	0.14	0.05	1.7
	0.1	2500	300	1.62	2.25	0.29	5.8	1.47	0.15	0.15	6.2
	0.1	4200	300	1.74	2.24	0.18	2.3	2.37	0.24	0.11	2.1
	0.1	4600	300	5.32	2.04	0.39	2.3	4.62	0.51	0.45	2.2
BeadBeater-8 (BioSpec)	0.1	3200	90	1.92	2.16	1.82	5.8	1.38	0.08	1.1	5.8
	0.1	3200	180	1.92	2.1	0.95	5.8	2.38	0.13	1.31	6.0
	0.1	3200	270	1.66	2.08	1.11	5.6	3.71	0.21	0.56	6.1
	0.5	3200	90	1.12	2.16	1.58	6.2	3.72	0.21	1.29	4.3
	0.5	3200	180	3.26	2.16	1.56	5.4	2.48	0.14	0.59	5.2
	0.5	3200	270	1.36	2.15	1.86	5.3	2.93	0.16	1.19	5.6
TissueLyser (Illumina)	0.5	1800	90	1.13	2.14	1.4	4.0	1.01	0.06	1.76	4.2
	0.5	1500	90	0.71	2.12	1.61	3.4	0.69	0.04	0.9	1.8
	0.5	1200	90	1.03	1.81	0.73	2.0	0.51	0.03	1.25	1.9
	0.5	1800	180	0.53	2.15	0.08	4.1	1.73	0.10	0.97	5.5
	0.5	1500	180	1.29	1.8	0.43	5.3	0.75	0.04	0.83	2.3
	0.5	1200	180	1.20	2.16	1.41	3.4	1.19	0.07	1.16	3.5

Table 3.3: Summary of RNA extractions performed with varying bead sizes, shaking speed and shaking time. Samples A and B are technical replicates of each other. RNA quantities, 260/280 ratio and 260/230 ratio were measured using the NanoDrop 2000c machine, while RIN values were obtained from the Bioanalyzer 2100 machine.

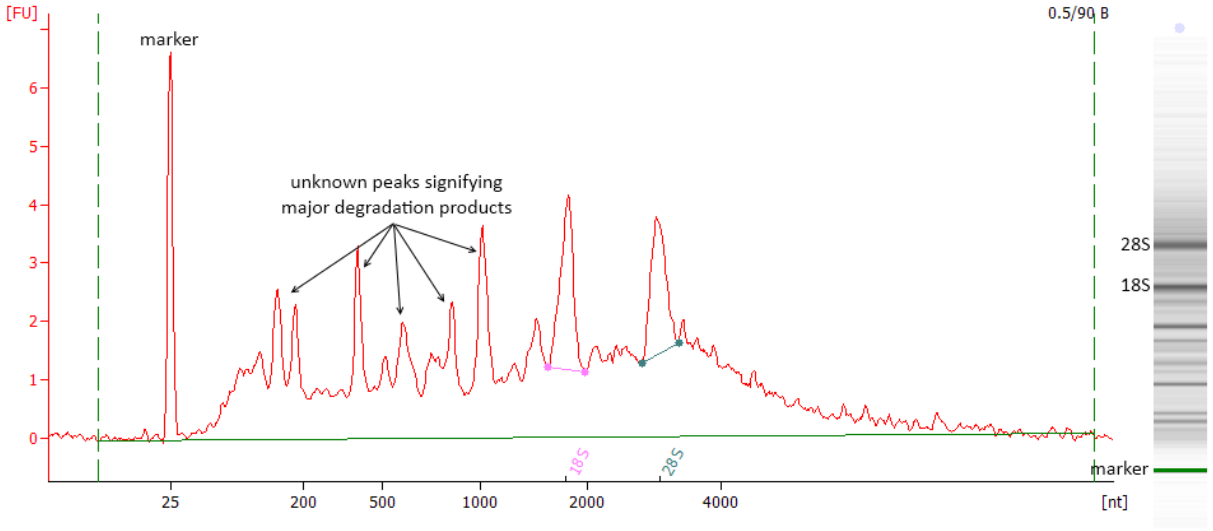


Figure 3.3: Electropherogram and gel representation of an RNA extraction done at a shaking speed of 3,200 rpm, shaking time of 90 seconds and bead size of 0.5 mm. The secondary peaks on the electropherogram correspond to the presence of secondary bands on the gel representation in between the marker and 18S band.

cells during the homogenisation step. However, as shown in Figures 3.4, 3.5 and 3.6, the band sizes were similar across samples. The same banding patterns were also observed (see Figure 3.7) when the samples were homogenised with a liquid nitrogen-cooled mortar and pestle, which discounts the possibility that bead-beating causes the RNA in the culture to degrade rapidly.

Following suggestions from Manuel Aranda (personal communication), as well as from the literature (Iida *et al.*, 2008; Rosic and Hoegh-Guldberg, 2010), we decided to switch to using exponentially-growing cultures instead of stationary phase ones. It is of note that there has not been any previous papers documenting the unsuitability of stationary phase cultures for RNA extraction. On the other hand, it has been observed that protein profiles from stationary and log phase cultures of *Symbiodinium sp.* are very similar to each other (Stochaj and Grossman, 1997), which implies that the expression profiles from both cultures would be similar to each other as well.

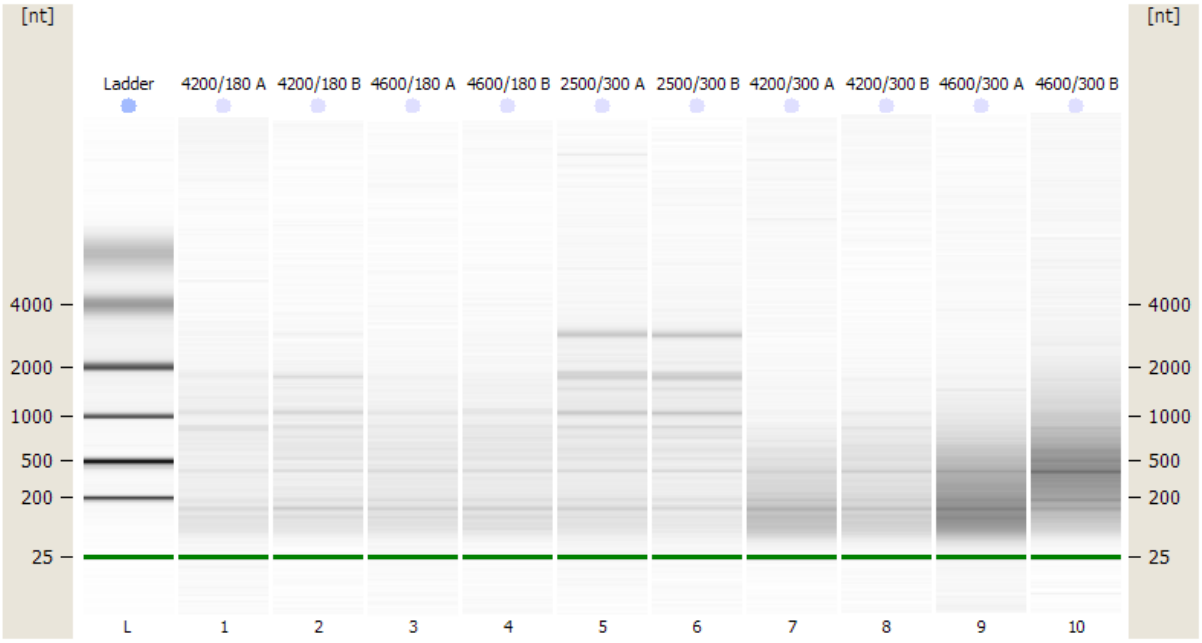


Figure 3.4: Gel representation of samples homogenised in the Mini BeadBeater-1 (Biospec). The labels at the top of the lanes indicate shaking speed and shaking time (e.g. 4200/180 means a shaking speed of 4,200 rpm with a shaking time of 180 seconds); "A" and "B" denote technical duplicates. Note the higher proportion of degraded RNA when the samples are homogenised for longer at a higher speed.

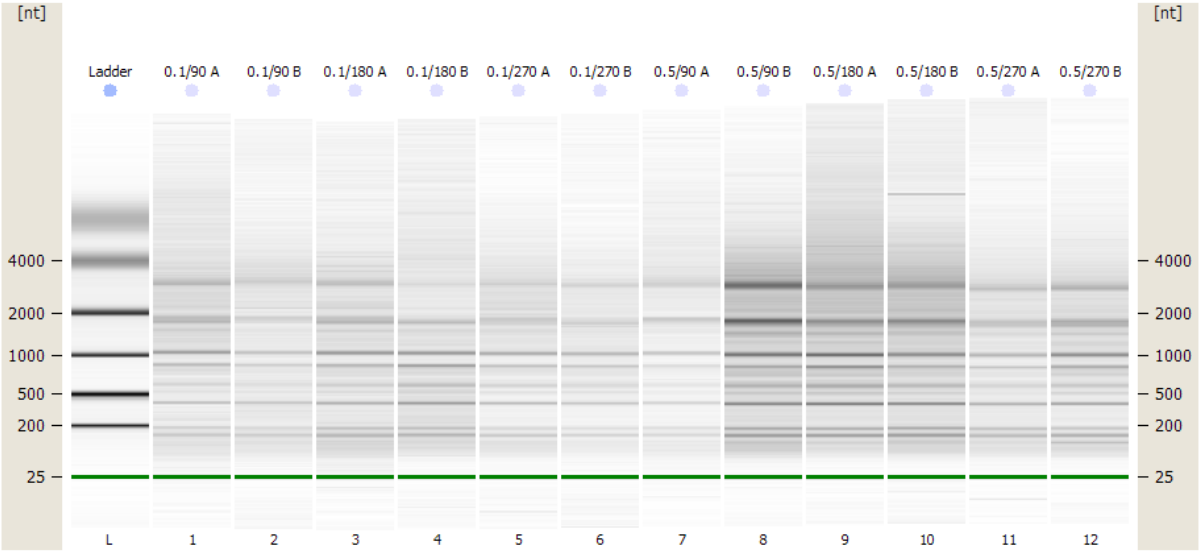


Figure 3.5: Gel representation of samples homogenised in the BeadBeater-8 (Biospec). The labels at the top of the lanes indicate bead size (in mm) and shaking time (e.g. 0.1/90 indicates the use of 0.1 mm beads with a shaking time of 90 seconds); "A" and "B" denote technical duplicates. Banding pattern is consistent across whole gel.

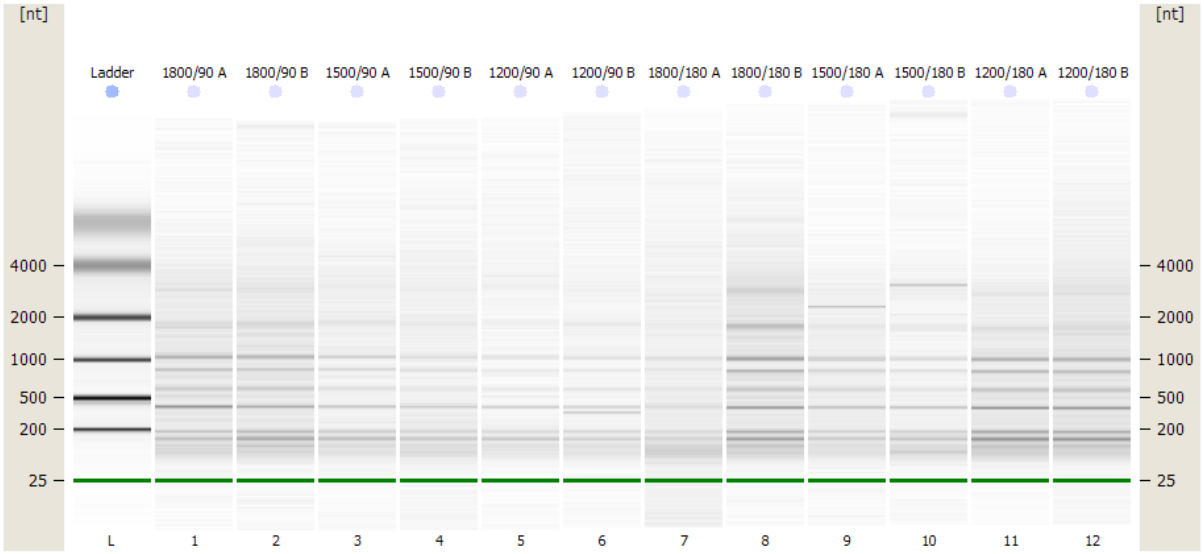


Figure 3.6: Gel representation of samples homogenised in the TissueLyzer (Illumina). The labels at the top of the lanes indicate shaking speed and shaking time (e.g. 1800/90 means a shaking speed of 1,800 rpm with a shaking time of 90 seconds); "A" and "B" denote technical duplicates. The lanes appear fainter on average due to the lower RNA yields.

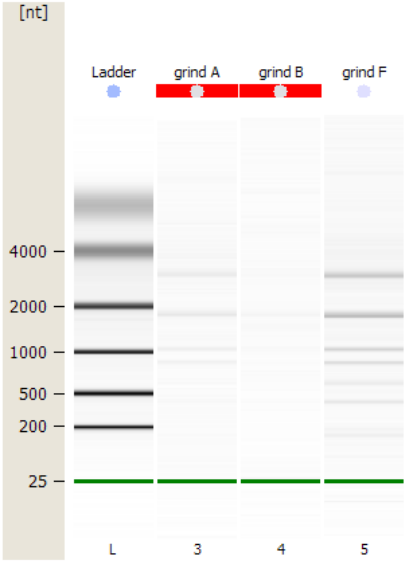


Figure 3.7: Gel representation of samples homogenised by grinding in a mortar and pestle. "A", "B" and "F" denote technical triplicates. Yields are much lower than that achieved through bead-beating techniques. Despite the gentler homogenisation method used, the secondary bands are still present.

3.3.2 RNA extraction from log phase cultures

Instead of repeating the previous experiments on the new exponential phase culture, we decided to compare three different RNA extraction protocols and contrast the quality of the resulting extracted RNA. For our desired downstream application (RNA sequencing), we reasoned that RNA quality is more of a concern than quantity. The quantity of the extracted RNA can be increased just by performing the extraction protocol on more cells, but high quality RNA might necessitate the use of a different protocol. This is compounded by the fact that Illumina sequencing only requires 1 µg of RNA to work, which is an amount we were able to regularly achieve in our RNA extractions (see Table 3.3).

The three protocols tested were:

1. Culture homogenised in liquid nitrogen-cooled mortar and pestle, RNA extracted using the mirVana kit (Ambion).

2. Culture homogenised in liquid nitrogen-cooled mortar and pestle, RNA extracted using the RNeasy Plant Mini Kit (Qiagen).

3. Culture homogenised in bead beater (shaking speed of 3,200 rpm, shaking time of 180 seconds and bead size of 0.1 mm), RNA extracted with phenol-chloroform method. The RNA extract was washed using the RNeasy Mini Kit (Qiagen).

The choice of homogenising cultures using a mortar and pestle was used by at least two groups in obtaining RNA of quality sufficient for RT-qPCR (Santiago-Vázquez *et al.*, 2006; Boldt *et al.*, 2009). The third protocol was carried out to determine the underlying reason behind the extensive RNA degradation observed in previous extractions. Results from these experiments is shown in Figures 3.8, 3.9 and Table 3.4.

From Figure 3.8 and 3.9, judging from the fainter bands and smaller peaks respectively, the RNA degradation observed in previous experiments can be attributed mostly to the use of stationary-phase samples.

From the results we obtained from these experiments, we chose mirVana as the protocol used to extract RNA from subsequent experiments. The benefit of using mirVana lies with

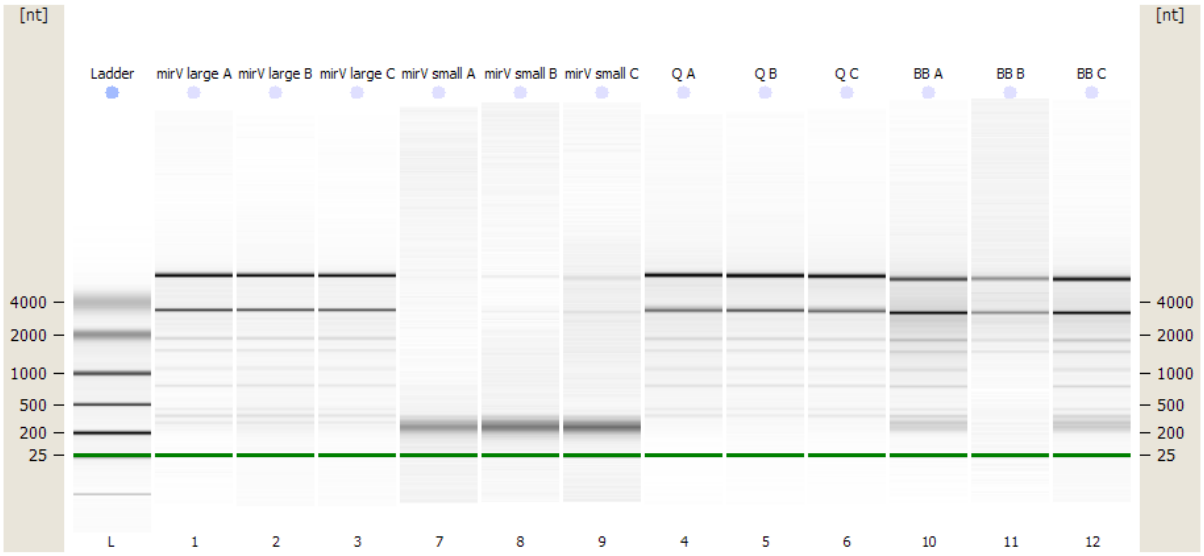


Figure 3.8: Gel representation of RNA extracted from exponentially-growing cultures using three different methods. All extraction methods were performed on the same starting amounts and from the same biological sample. A, B and C represent technical triplicates. "mirV large" represents the large RNA fraction obtained using mirVana (Ambion) while "mirV small" the small RNA fraction from mirVana; "Q" represents RNA obtained using the RNeasy Plant Mini Kit (Qiagen); "BB" represents RNA obtained from homogenising samples in a bead-beater.

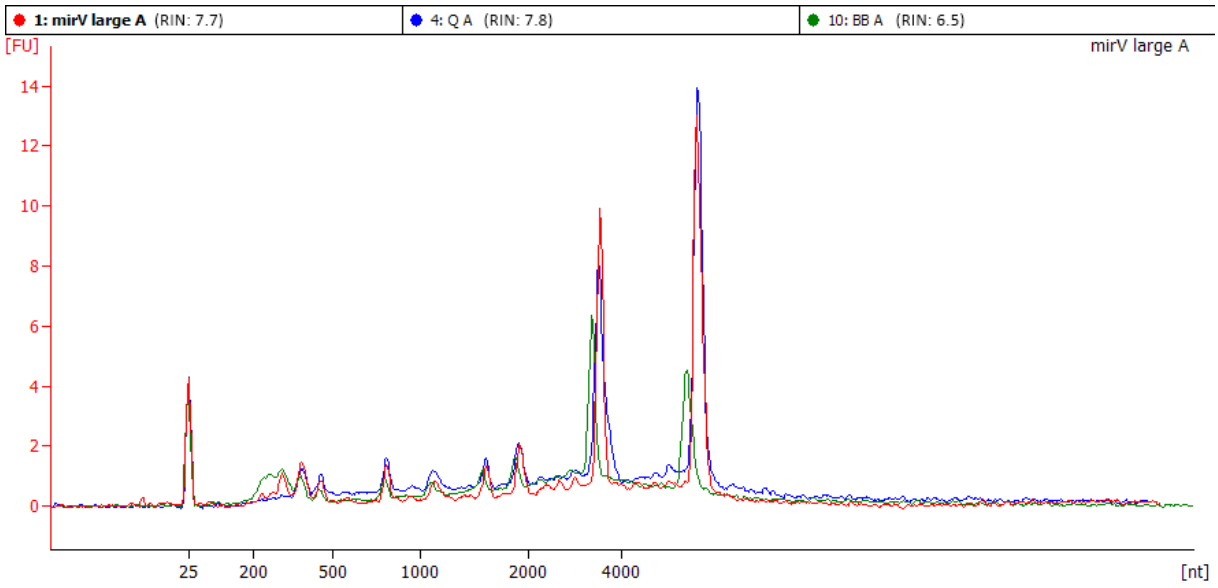


Figure 3.9: Electropherogram contrasting RNA qualities obtained from three different extraction methods. "mirV large" represents the large RNA fraction obtained using mirVana (Ambion); "Q" represents RNA obtained using the RNeasy Plant Mini Kit (Qiagen); "BB" represents RNA obtained from homogenising samples in a bead-beater.

Protocol	Replicate	Quantity / µg	260/280 ratio	260/230 ratio	RIN value
mirV large mirVana (Ambion)	A	3.46	2.14	1.86	7.7
	B	2.98	2.14	1.96	5.4
	C	4.27	2.14	2.01	5.4
Q RNeasy Plant Mini Kit (Qiagen)	A	2.48	2.12	2.22	7.8
	B	2.29	2.14	0.34	7.8
	C	2.66	2.12	2.22	8
BB Bead-beater, phenol-chloroform extraction	A	1.49	2.07	0.67	6.5
	B	0.53	2.11	0.39	8
	C	1.38	2.11	1.71	7.1

Table 3.4: Summary of RNA extracted using three different methods. Samples A, B and C are technical replicates of each other. RNA quantities, 260/280 ratio and 260/230 ratio were obtained from the NanoDrop 2000c machine, while RIN values are calculated from the Bioanalyzer 2100 machine.

the fact that we could separate an RNA fraction which was enriched in small RNA ($<$ 200 bp) from the total RNA, as seen in Figure 3.8. We intended to sequence the small RNA fraction independently to study the small RNA-ome of *Symbiodinium sp.* Although RNeasy Plant Mini Kit (Qiagen) produced RNA of better qualities than mirVana (see Table 3.4), RNeasy kits remove most of the small RNA fraction (including tRNAs) during the extraction procedure, which makes it unsuitable for small RNA-ome studies.

A closer examination of Figure 3.9 supports the conjecture that bead-beating techniques cause the partial degradation of rRNA, especially 28S rRNA ("BB A", green line on electropherogram). While the difference in absolute peak heights between the green and other two lines can be attributed to the lower RNA concentration of "BB A" (refer to Table 3.4), the 28S:18S rRNA ratio have dropped significantly for cultures were homogenised in a bead-beater, which indicates degradation of 28S rRNA.

3.3.3 RNA extractions performed on cultures subjected to different stresses

Based on the results from the previous section, cells from nine separate exponentially-growing cultures subjected to different environmental stresses was homogenised in a liquid nitrogen-cooled mortar and pestle, and RNA extracted using mirVana (full protocol in Appendix).

The results from these extractions are summarised in Table 3.5.

3.4 Conclusion

As RNA-Seq was used to survey the small RNAome of *Symbiodinium sp.* (discussed in the next chapter) and also to profile transcriptional changes of *Symbiodinium* cultures subjected to different stresses, we were interested in methods that could extract high-quality RNA from our cultures to obtain better sequencing data.

Due to the presence of a strong, chemically resistant cell wall, typical methods used to extract RNA from metazoan tissues were ineffective on *Symbiodinium sp.* cells. The paucity of literature regarding optimised methods at extracting RNA from these cells spurred us to optimise two factors that affect the eventual quality of the extracted RNA.

Firstly, the choice of homogenisation method has to strike a balance between being harsh enough to break apart the cell wall using mechanical means, and gentle enough to prevent RNA degradation in the cells. Although one group has had success in homogenisation via bead-beating methods, our efforts using the same method resulted in moderately degraded RNA. We eventually succeeded in producing high-quality RNA by homogenising the cells in a liquid nitrogen-cooled mortar and pestle.

Secondly, the growth phase of the culture has an effect on the extracted RNA quality as well. The presence of degraded RNA was higher in cultures from stationary phase than log phase. This observation has not been documented in the literature.

Based on the optimised RNA extraction protocol, we were able to extract RNA of high quality (having a RIN value of > 7.8) from our *Symbiodinium* cultures. The analysis of the sequence data from these extracts will be expounded upon in the next chapter.

Condition	Replicate	Large RNA fraction				Small RNA fraction			
		Amount / µg	**260/280**	**260/230**	**RIN**	**Amount / µg**	**260/280**	**260/230**	**RIN**
Cold stress (4 °C for 4 hours)	A	1.15	2.05	0.61	2.6	0.55	1.8	0.07	-
	B	0.96	1.87	1.43	6.6	0.59	1.99	0.24	-
Low temperature (16 °C for 4 hours)	A	6.3	2.13	1.98	7.3	1.44	1.98	0.16	-
	B	2.79	2.1	0.66	7.8	1.1	2.03	0.08	-
Heat stress (36 °C for 4 hours)	A	1.36	2.02	0.1	7.5	0.81	1.95	0.06	-
	B	1.6	2.08	1.17	8.3	0.43	1.99	0.29	2.5
Hypertonic stress (60 g/l NaCl for 4 h)	A	2.28	2.1	1.76	8	0.64	1.98	0.18	-
	B	1.95	2.11	1.61	7.7	0.7	1.95	0.18	-
Hypotonic stress (20 g/l NaCl for 4 h)	A	3.28	2.13	1.59	7.6	1.22	1.74	0.32	2.5
	B	3.81	2.18	2.04	7.9	0.75	2.38	2.47	2.5
Dark stress (+12 h of darkness)	A	2.22	2.11	0.92	7.8	0.45	2.04	2.2	2.6
	B	2.32	2.17	0.58	8.3	0.47	1.96	1.34	2.5
Heat stress (34 °C for 12 hours)	A	4.92	2.15	1.86	8.1	1.25	2.05	0.07	2.6
	B	3.51	2.17	1.36	8	1.32	2.04	0.96	-
Regular night (harvest at midnight)	A	2.45	2.09	1.9	8.1	0.62	1.89	1.45	4.6
	B	2.99	2.1	2.22	8.1	0.65	1.99	0.22	2.6
Regular day (cells harvested at noon)	A	3.12	2.14	1.86	8	0.82	2.06	2.1	2.6
	B	2.69	2.14	1.96	8.1	1.07	1.96	0.24	2.5
	C	3.85	2.14	2.01	7.9	1.05	1.95	1.69	2.5

Table 3.5: Summary of RNA extractions performed on nine cultures subjected to different stresses. Samples A, B and C are technical replicates of each other. RNA quantities, 260/280 ratio and 260/230 ratio were measured using the NanoDrop 2000c machine, while RIN values were obtained from the Bioanalyzer 2100 machine.

Chapter 4

Study of small RNAome for *Symbiodinium sp.* and *Stylophora pistillata*

4.1 Introduction

Stylophora pistillata, also known as "hood coral" or "smooth cauliflower coral", is commonly found in the tropical seas of the world. As tropical seas tend to be nutrient-poor, the endosymbiotic relationship between the coral host and unicellular algae symbionts is important for the survival of the coral. Of the many symbionts that could associate with the coral, the most typical one would be the photosynthetic dinoflagellate of the genus *Symbiodinium sp.* (Weston et al., *2012*).

Due to the large evolutionary gap between these two organisms (see Figure 4.1), the evolutionary histories for both organisms will be outlined separately, followed by an outline for the various motivations that underlie the study of these two organisms. This section rounds off with a survey of small RNA studies in related organisms, which will serve as a basis for functional comparisons between our organisms of interest and those related organisms.

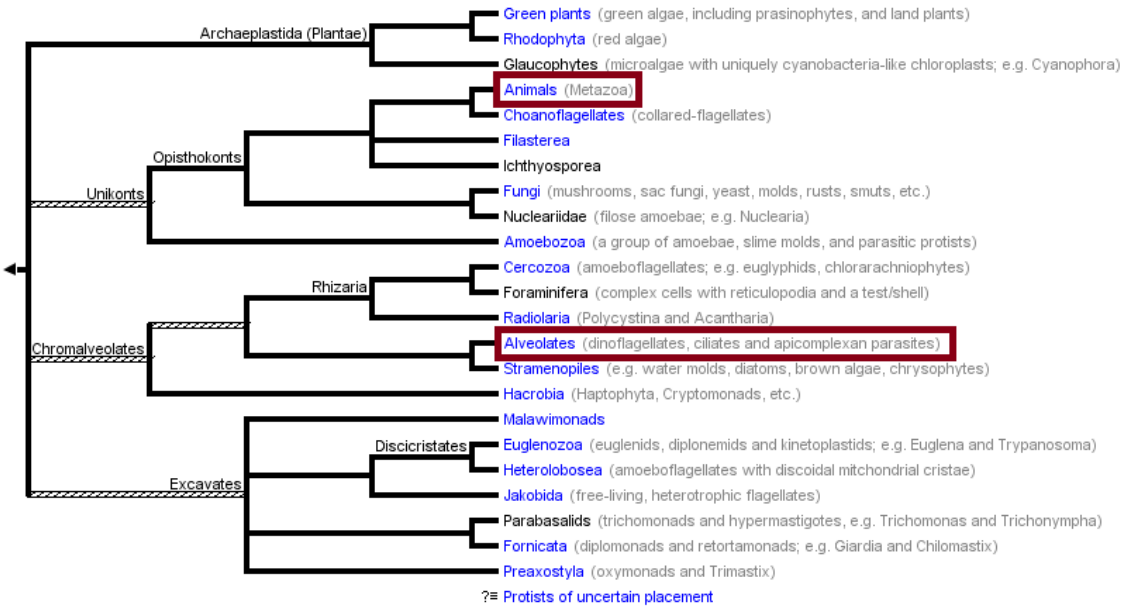


Figure 4.1: A tree showing the phylogenetic relationships between eukaryotes. The large evolutionary distance between the metazoan *S. pistillata* and the alveolate *Symbiodinium sp.* is apparent from this tree (adapted from Tree of Life web project, `http://tolweb.org/Eukaryotes/3`).

4.1.1 Evolutionary histories

4.1.1.1 *Symbiodinium sp.*

The genus *Symbiodinium*, also colloquially known as "zooxanthellae" (Blank and Trench, 1986), consists of photosynthetic dinoflagellates that are able to form symbiotic relationships with other marine invertebrates and protists, such as cnidarians (e.g. corals, sea anemones and jellyfish), flatworms, molluscs, sponges and foramniferans (Stat *et al.*, 2006). It is important to note that this symbiotic relationship is not specific: *Symbiodinium* is not the only genus of dinoflagellates that are endosymbionts of those invertebrate and protist hosts. This is not surprising, considering that there are ~2,000 species of dinoflagellates, abundant in both marine and freshwater environments, and half of which are photosynthetic (Taylor *et al.*, 2008). However, among the eight distinct genera that are able to forge symbiotic relationships with their hosts, *Symbiodinium* is the one that attracted the most scientific inquiry (Baker, 2003).

In the past, due to the fairly uniform *in symbio* appearance of *Symbiodinium* cells under

a light microscope — they appear as brown coccoid cells that are about 5–15 µm in diameter — *Symbiodinium* was mistakenly thought to be one single species (*Symbiodinium microadriaticum*). Questions regarding the homogeneity of *Symbiodinium* started to arise in the 1980s, as newer behavioural, physiological and ultrastructural evidence hinted at greater diversity within *Symbiodinium* (Weis *et al.*, 2008). With the advent and widespread use of more sophisticated technologies in biology, studies have revealed significant differences (e.g. chromosome numbers, chloroplast arrangement, cell physiology and metabolic processes) within the *Symbiodinium* genus (Stat *et al.*, 2006).

Although *Symbiodinium* is now widely accepted as a group of highly diverse dinoflagellates, taxonomic efforts at classifying and naming distinct species within this genus is hampered by the lack of any observed sexual reproduction within *Symbiodinium*. This makes it impossible to categorise distinct species according to the biological species concept. Despite this drawback, there are at least eleven named *Symbiodinium* species. Four species (*S. microadriaticum*, *S. pilosum*, *S. kawagutii* and *S. goreaui*) have been formally defined according to the morphological species concept, but the lack of formal descriptions for the other named *Symbiodinium* species (e.g. *S. corculorum*, *S. pulchrorum*) makes it possible that these names are synonymous with existing ones (Baker, 2003). More recently, the reliability of using morphology in delineating species boundaries has come into question due to the phenotypic plasticity demonstrated by *Symbiodinium* cells. The morphology of these cells have been shown to vary depending on culture conditions, initial growth conditions in hospite (when the *Symbiodinium* was still associated with the host), and the intensity of light from the surroundings. Also, as different *Symbiodinium* species have different successes in culture, the relative abundance of a certain species in culture will not be an accurate reflection of its *in symbio* abundance (Stat *et al.*, 2006), which makes the ecological species concept unreliable at defining species boundaries as well.

As such, our current understanding of the phylogenies within *Symbiodinium* is largely based on the molecular efforts throughout the past 30 years. Most studies have focused on *Symbiodinium* rDNA regions (18S, 28S and internal transcribed spacer (ITS) regions);

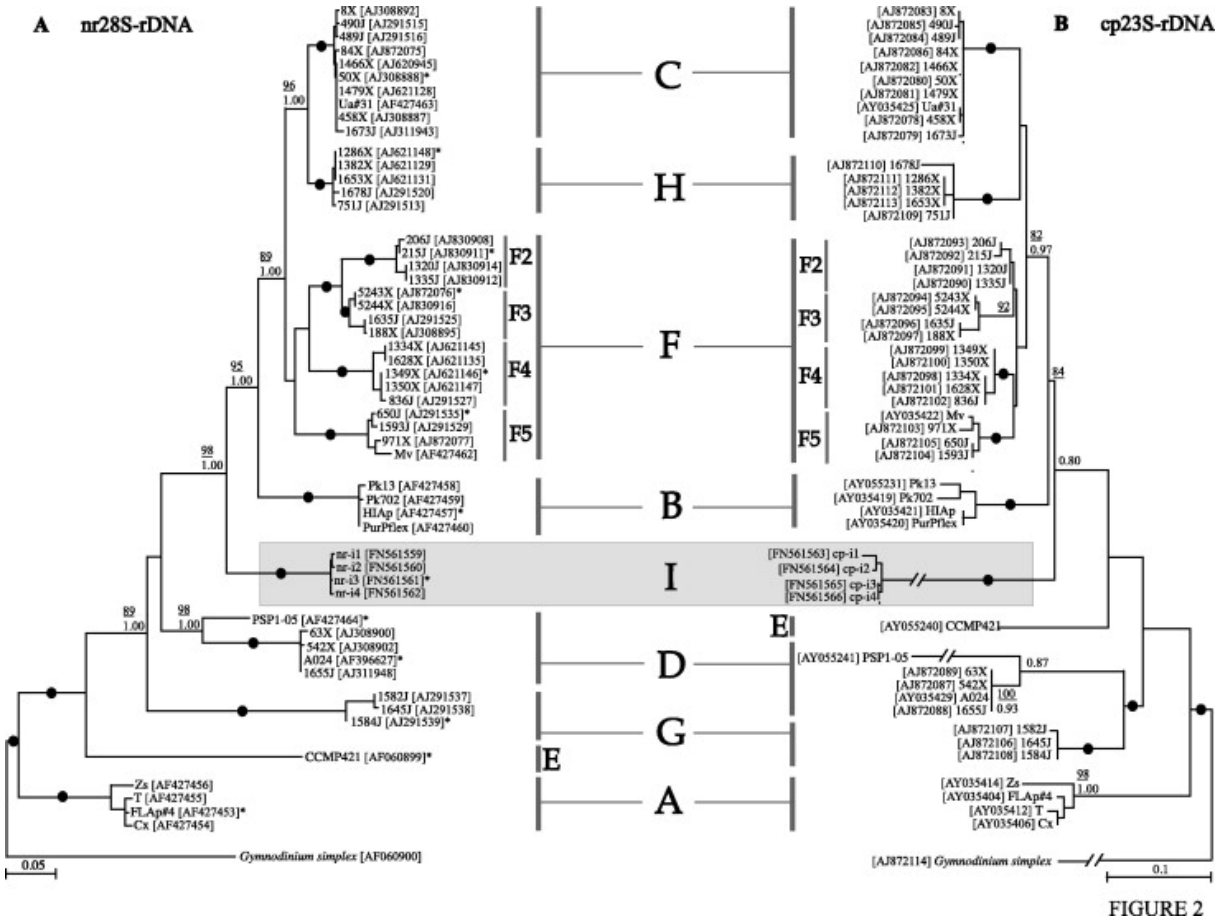


Figure 4.2: Maximum likelihood phylogram for the nine *Symbiodinium* clades. Phylogram constructed based on nuclear 28S rRNA and chloroplastic 23S rRNA sequences (labelled as nc28S-rDNA and cp23S-rDNA respectively) (Pochon and Gates, 2010).

while other studies either looked at other genes (e.g. allozyme, chloroplast rDNA and *psbA*, mitochondrial *cox1*) or used other PCR-based techniques (e.g. RAPD, microsatellite analysis, DNA fingerprinting) (Stat *et al.*, 2006).

Currently, there are nine named clades within *Symbiodinium* (see Pochon *et al.* (2004); Stat *et al.* (2006) for a review of clades A–H; Pochon and Gates (2010) for the discovery of clade I). A molecular phylogeny of these nine clades is detailed in Figure 4.2. *Symbiodinium* in clades A–D are usually associated with metazoan hosts, while clades C, D, F, G, H and I tend to associate with benthic foraminiferans (Pochon and Gates, 2010). Among clades, the greatest diversity is seen in clade C — not just in terms of rDNA molecular types, but also the number of hosts that could form mutualistic associations with *Symbiodinium* (LaJeunesse, 2005).

It is not uncommon for a single host to harbour *Symbiodinium* from multiple clades, and interestingly, some hosts are able to cope with stresses by altering the composition of *Symbiodinium* in the host cells. For example, *Acropora* mainly associates with *Symbiodinium* from clade C in cooler seas, but in hotter regions, the *in symbio* compositions shift to favour the more thermally-tolerant clade D (Oliver and Palumbi, 2009). This composition shift is not devoid of tradeoffs: *Acropora* with a majority of clade D symbionts grow ∼30% slower than those with clade C symbionts (Jones and Berkelmans, 2010). The complex interactions between *Symbiodinium* and its host, as well as the changes in host phenotypes due to different symbiont compositions, are actively being studied.

4.1.1.2 *Stylophora pistillata*

Stylophora pistillata is a reef-building coral (order Scleractinia, family Pocilloporidae), first catalogued as a species by Eugen Esper in 1797. The coral is branched, with stout branches in shallow water but gradually becomes finer in deeper waters. In deeper regions, the branches are usually arranged in a manner that maximises light capture. This coral can be found in a wide range of habitats, and takes on a wide range of colours: cream, pink, blue or green (Veron, 2000).

The fossil record of scleractinians dates back to the Triassic (∼250 million years ago) (Romano and Cairns, 2000). The early Scleractinians were not reef-builders — there is a time gap of 20–25 million years separating the earliest Triassic corals and the earliest known coral reefs (the ancestors of the reef-building Pocilloporidae originated in the late Triassic). The scleractinians continued to flourish into the Jurassic time period (200–145 mya), and it is thought that the Late Jurassic had the all-time highest coral diversity. Many of the extant coral families have origins that date back to the Jurassic time period (a family tree is shown in Figure 4.3) (Veron, 2000).

Past understanding of the evolutionary history of Scleractinians has been mainly guided by the skeletal characteristics of the corals. Such efforts date all the way back to the mid-19th century. However, the emergence of molecular techniques in constructing phylogenetic

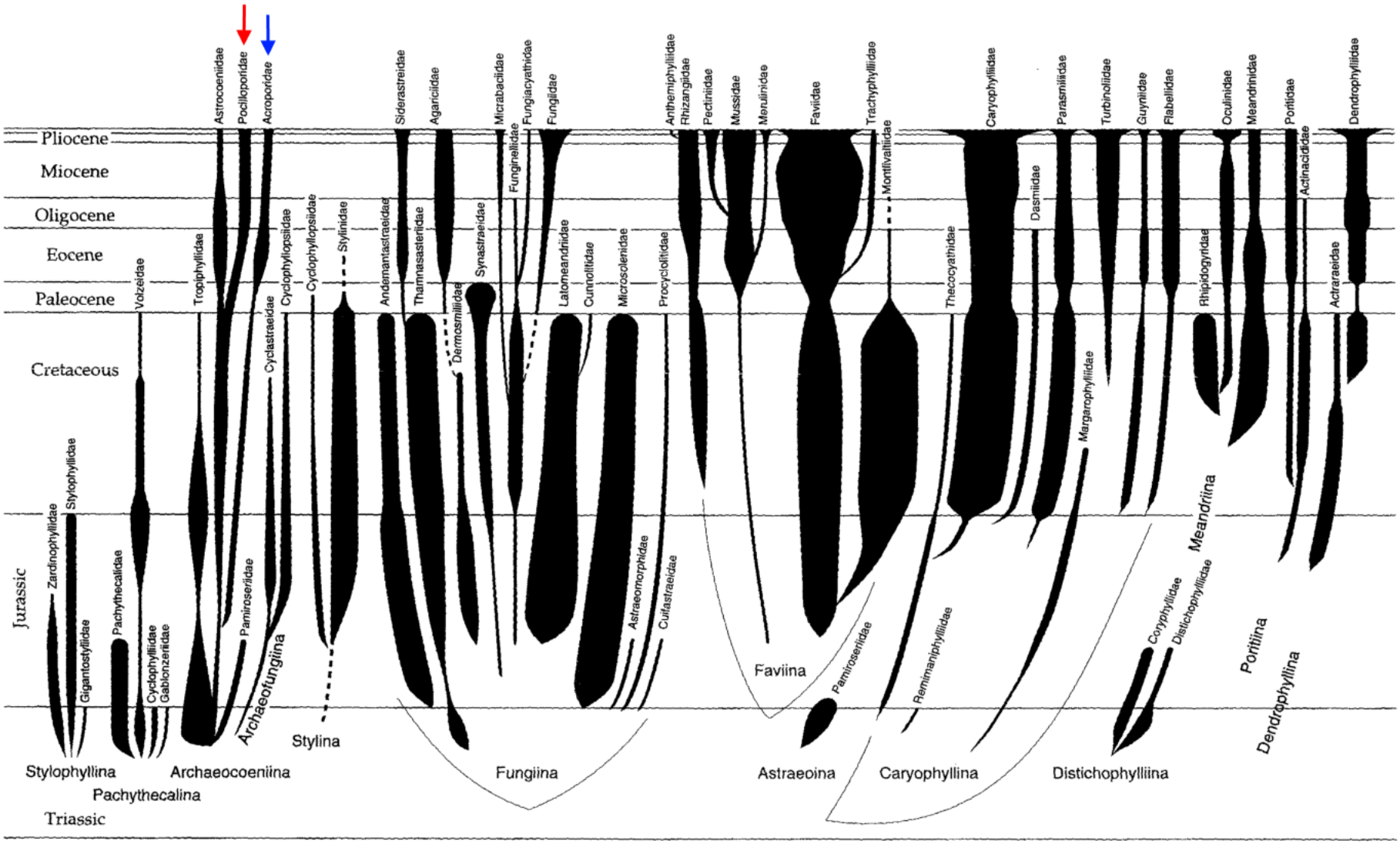


Figure 4.3: Family tree of major families in Scleractinians. Width of bars indicates estimated global abundance, not diversity within families. Pocilloporidae, marked with the red arrow, is the family which *S. pistillata* belongs to; Acroporidae, marked with the blue arrow, indicates the family that has a member with a sequenced genome (*Acropora digitifera*). Illustration adapted and modified from Veron *et al.* (1996).

relationships between extant coral families that are independent of skeletal data has resulted in several disagreements with the traditional morphology-based classification (Stolarski and Roniewicz, 2001). For instance, the phylogenetic analysis based on mitochondrial 16S rDNA data by Romano and Palumbi (1996) provides support for grouping of extant corals into morphologically-based families, but not that of traditional suborders. Also, it was proposed that Scleractinians had a divergence into two different clades (termed "robust corals" and "complex corals" respectively) ∼240 million years ago, prior to the appearance of the scleractinian skeleton. There is no clear-cut criterion that differentiates the morphology of members from both clades. However, generally speaking, "robust corals" have a more heavily calcified skeleton and reproduce via intratentacular budding; "complex corals" have a less heavily calcified skeleton and reproduce via extratentacular budding.

In a separate study, Veron *et al.* (1996) conducted a phylogenetic analysis based on nuclear 28S rDNA data. In agreement with the previously-mentioned analysis, theirs supports grouping extant corals into its traditional families too, but did not agree with the groupings of families into suborders as proposed by Romano and Palumbi (1996). Interestingly, the proposed early divergence of Scleractinians into two clades did also find support in the 28S rDNA data.

The discrepancies between molecular and morphological data, and the resulting effects on the classification of Scleractinians, have yet to be fully discussed and summarised in the form of a Treatise (Stolarski and Roniewicz, 2001). Phylogeny in Scleractinians is far from being certain, and remains under active study.

4.1.2 Considerations behind choice of organisms for sequencing

4.1.2.1 *Symbiodinium sp.*

The interesting biology of *Symbiodinium* notwithstanding, there are multiple other reasons why *Symbiodinium* is currently considered the best candidate for being the first dinoflagellate genome to be fully sequenced.

One of the remarkable biological oddities of dinoflagellates is that they possess huge

genomes, despite the relative simplicity of the cell. Measurements of DNA content in dinoflagellates using various methods have produced estimates that range from 3 pg cell^{-1} (*Symbiodinium*) to 280 pg cell^{-1} (*Prorocentrum micans*) (Veldhuis *et al.*, 1997; LaJeunesse *et al.*, 2005), which is approximately 1–40 times the size of a human diploid genome. As *Symbiodinium* has one of the smallest genomes among dinoflagellates, it serves as a good entry point for initial sequencing efforts. It is likely that there are dinoflagellate genomes much smaller than *Symbiodinium* — Lin (2006) makes a case for picoplanktonic dinoflagellates, which are an order smaller in physical size than *Symbiodinium*, being likely to have genomes smaller than *Symbiodinium* as well. However, the picoplanktonic dinoflagellates have yet to have well-documented mutualistic or parasitic relationship with other hosts (Callieri *et al.*, 2007). Other non-*Symbiodinium* dinoflagellates that are able to form symbiotic relationships with cnidarians (e.g. *Amphidinium carterae*, *Alexandrium tamarense*) have been shown to have genomes ∼2–30 times that of *Symbiodinium* (Trench, 1997; LaJeunesse *et al.*, 2005).

It has also been shown that *Symbiodinium* is amenable to genetic manipulation — plasmids containing selectable markers (*amp*, *npt II*, *hpt II*) and a reporter gene (GUS) were successfully introduced via silicon carbide (SiCa) whiskers into *Symbiodinium* cells. The resulting transformation efficiency (5–24 transformants per 10^7 cells) is comparable to ballistic methods on diatoms and SiCa methods on *Chlamydomonas* (ten Lohuis and Miller, 1998). This study, however, remains as the only reported attempt at transforming *Symbiodinium* (Weis *et al.*, 2008) — further efforts in developing or refining existing techniques is needed if *Symbiodinium* were to become the model organism for studying dinoflagellates.

Another point in favour of sequencing the genome of *Symbiodinium* is that the genome should be similar to other dinoflagellates. There are some concerns that the long-term mutualistic association with other marine invertebrates might have substantially modified the genome of *Symbiodinium*, but this concern seems unlikely — many *Symbiodinium* species are able to survive for months *in vitro* as independent, free-living dinoflagellates, indicating that the full complement of housekeeping and other vital genes in common with other dinoflagellates are still present in *Symbiodinium* (LaJeunesse *et al.*, 2005).

4.1.2.2 *Stylophora pistillata*

Scientific curiosity of corals in general has recently intensified, following observations corals are dying at an alarming rate worldwide — for instance, in the Caribbean, Hughes (1994) reports that coral cover (the extent of sea floor occupied by corals) has declined from over 50% in the 1970s to less than 5% in the 1990s; in the Indo-Pacific region, home to 75% of the world's coral reefs, Bruno and Selig (2007) estimated that coral cover declined ∼1% annually in the past 20 years, and ∼2% between 1997–2003. This trend is worrying, as coral reefs are the "rainforests of the sea", supporting more marine biodiversity per unit area than other marine habitats. Despite only covering less than 0.2% of the ocean's surface, corals are the home for an estimated 35% of all marine species (Knowlton *et al.*, 2010). There are many reasons behind the destruction of coral reefs, which include, but are not limited to, overfishing (Hughes, 1994); pollution (Pastorok and Bilyard, 1985; Bak, 1987); disease (Green and Bruckner, 2000) and accelerated warming and acidification of oceans (Hughes *et al.*, 2003).

Past research on *S. pistillata* dates well back into the 1970s. Within the coral community, proposals for having a model system in corals have only recently gained traction. A few years ago, Weis *et al.* (2008) argued for the need of model systems in corals, so that researches can coordinate and crystallise scientific efforts in understanding coral cell biology, as well as refine techniques used in studying the corals. *Stylophora pistillata* is one of the four proposed model organisms, the other three being *Aiptasia sp.*, *Acropora millepora* and *Acropora palmata.* More recently, an Australian initiative called "ReFuGe 20/20" (short for "Reef Future Genomics") aims to sequence ten coral genomes in the near future, with the stated goals of improving our understanding of coral physiology and reef management strategies based on the data made available by the initiative.

We chose to work on *S. pistillata* for many reasons. We are interested in studying the stable and long-term symbiotic relationship of this coral and the photosynthetic *Symbiodinium* dinoflagellate at the molecular level, for example: the identities of the protein transporters that translocate photosynthetic products from the symbiont to the host, and inorganic nutri-

ents in the other direction; gene transfer, if any, between host and symbiont (Shinzato *et al.* (2011) notes that there is no evidence of gene transfer in *Acropora*); and genes involved in the exocytosis of the symbiont when the host is under prolonged stress. Also, the genome will reveal the molecular underpinnings that belie coral health and responses to environmental stresses, which could be incorporated into strategies that reduce, and hopefully reverse, the worldwide decline in coral reef cover.

Also, *S. pistillata* is very common in the world's seas, both in terms of area (see Figure 4.4) and abundance. It is the most abundant and widespread reef-building coral in the Gulf of Eilat, Northern Red Sea (Weis *et al.*, 2008). Its distribution across large differences in temperature ranges, ranging from the Red Sea to the Southern Ocean, makes it a good candidate for studying adaptability and thermal tolerance in corals.

Figure 4.4: Worldwide distribution of *S. pistillata*. Boundaries of the highlighted regions are defined according to the guidelines in the "Coral Geographic" (unpublished), which divides the world's coral regions into 141 named ecoregions (Veron, 2000).

Other advantages in using *S. pistillata* is its ease in being cultured *in aquaria* (Weis *et al.*, 2008) — in fact, it has been cultured successfully in the Centre Scientifique de Monaco for over 20 years, following the discovery of an efficient method of culturing coral in aquariums (Jaubert, 1989). It has also been reported to recover fairly easily from experimental handling and breakages (Koren *et al.*, 2008).

As of the time of writing, there is one complete coral genome in literature: that of *Acropora digitifera* (Shinzato *et al.*, 2011). *Stylophora* differs considerably from *Acropora* — *Acropora* is a "complex coral", while *Stylophora* is a "robust coral". Both coral groups diverged ∼240

million years ago (Romano and Palumbi, 1996; Romano and Cairns, 2000). Also, both corals differ in terms of breeding strategies. *Stylophora* is a brooder coral (releases fewer, fertilised planula larva in a monthly cycle), while *Acropora* is a broadcast spawner (releases large amounts of eggs and sperm yearly) (Harrison and Wallace, 1990).

There are two other complete cnidarian genomes in literature: hydra (Chapman *et al.*, 2010) and sea anemone (Putnam *et al.*, 2007), but the large evolutionary distance between these two organisms and corals (e.g. ∼500 million years separate corals and sea anemone (Shinzato *et al.*, 2011)) makes the genomes ill-suited for the in-depth understanding of coral biology and physiology.

4.1.3 Survey of RNAi machinery and small RNAs in related organisms

4.1.3.1 *Symbiodinium sp.*

As of the time of writing, the underlying molecular machinery involved in the biogenesis and function of small RNAs (sRNAs) in dinoflagellates is poorly understood. While there are several papers (e.g. Leggat *et al.* (2007) and Bayer *et al.* (2012)) that describe the transcriptomic landscape of *Symbiodinium* from different clades, the two key proteins for minimal RNAi function — Dicer and Argonaute — has yet to be identified any *Symbiodinium*. Although Leggat *et al.* (2007) identified 12 contigs that bear similarity to known RNA processing and modification proteins, neither inferred functions nor sequences were provided for these contigs.

While there has not been any literature regarding the identification of (sRNAs) in any dinoflagellates, among other distantly-related protists (kingdom Chromalveolata), functional sRNAs have been identified in two other organisms: *Tetrahymena thermophila* and *Paramecium tetraurelia*.

In *T. thermophila*, there are three distinct classes of sRNAs: a ∼30–35 nt class that accumulate during starvation; a ∼27–30 nt class thought to be involved in the developmentally regulated DNA elimination; and a ∼23–24 nt class that might serve as guide strands for RNA

cleavage. Lee and Collins (2006) identified 272 sRNAs across the three classes mentioned above, but only one of these small RNAs is present in either of the *Symbiodinium* libraries at moderate abundances (511 reads, 0.0004% of library size).

In *P. tetraurelia*, the accumulation of a class of ∼22–23 nt sRNAs correlates with the homology-dependent silencing of maternal genes in developing germ line cells. Interestingly, the increase in sRNA levels can also be artificially induced by feeding cells with bacteria that contain long dsRNA precursors of those sRNAs, which strongly suggests the conservation of the RNAi pathway in Paramecia (Galvani and Sperling, 2002; Garnier *et al.*, 2004).

Judging from our short read data (data not shown), there is virtually no evidence supporting the conservation of the functional sRNA classes identified from these two protists in *Symbiodinium sp.*

4.1.3.2 *Stylophora pistillata*

While small RNAs are very well studied among metazoans, to date, none has been identified in any Scleractinian (there was no mention of small RNAs in the *Acropora digitifera* genome paper).

Branching further out, functional sRNA families such as miRNAs and piRNAs have been discovered in the sea anemone *Nematostella vectensis*, which shares the same class as the Scleractinians (class Anthozoa). Dicer and Argonaute, the two core proteins present in all organisms with a functional RNAi pathway, has been shown to be present in *N. vectensis* (Grimson *et al.*, 2008). In addition to these two core proteins, piRNAs and Piwi proteins are present as well. Based on the presence of these proteins, as well as the availability of short read sequence data, 40 miRNAs were predicted for *N. vectensis*. Among these predictions, only one of them is a near match to a known miRNA (as catalogued in miRBase): miR-100. *N. vectensis* miR-100 is offset by one nucleotide compared to other bilaterian miR-100s, as shown in Figure 4.5. All of the other novel predictions have yet to have functions assigned to them.

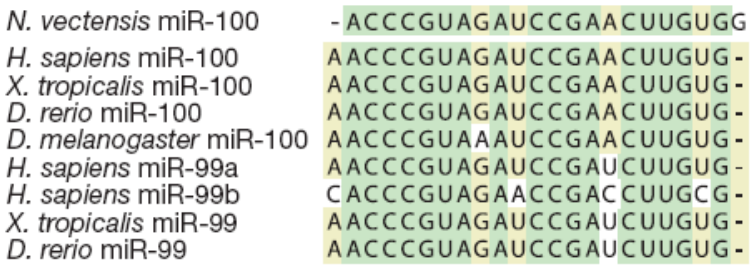


Figure 4.5: Comparison of *N. vectensis* mature miR-100 sequence against related miRNAs from other model organisms. The one-nucleotide offset in the miRNA sequence is expected to significantly change the miRNA-mRNA targeting, as target recognition is dependent on the primary sequence of nucleotides 2–7 (seed region). Illustration from Grimson *et al.* (2008).

4.2 Materials and methods

4.2.1 Sequence data used

Currently, the deep sequencing of the genomes, transcriptomes and proteomes of both *Symbiodinium sp.* (strain CCMP2467) and *S. pistillata* (cultivated *in aquaria* at Centre Scientifique de Monaco) is still ongoing. This project is a collaboration between the Voolstra Lab in KAUST (Saudi Arabia) and the Micklem Lab in Cambridge (UK), of which I am a member of the latter. I am primarily involved in the creation and analysis of the small RNA datasets from both organisms, but in this chapter, I have also used genomic, transcriptomic and proteomic data generated by others in the project.

A fuller description of the sequence data used in this chapter can be found in Tables 4.1 and 4.2.

Dataset	Filename	Remarks
Genome	symb_genome_v4.fa	Genome size estimated to be ∼936m bp.
Transcriptome	symb_genome_v4_transcripts_all.fa	∼52,000 predicted transcripts.
Proteome	symb_genome_v4_proteins_datafreeze.fa	∼36,000 putative sequences.
Small RNAome	s_7_sequence.symb.fastq	∼31.1m short reads.
	symbiodinium_all_indiv_conds.fastq (amalgamation of 9 separate files)	∼137m short reads.

Table 4.1: List of *Symbiodinium sp.* datasets used. The datasets contain preliminary sequence data for the genome, transcriptome, proteome and small RNAome, as of approximately July 2012.

Dataset	Filename	Remarks
Genome	styl_genome_v4.fa	Genome size estimated to be ∼793m bp.
Transcriptome	styl_genome_v4_transcripts_beta_anno.fa	Currently ∼83,000 predicted transcript and protein sequences, many of of which are duplicated. Expected ∼25,000 sequences for both.
Proteome	styl_genome_v4_proteins_beta_anno.fa	
Small RNAome	s_8_sequence.coral.fastq	∼30.5m short reads.

Table 4.2: List of *S. pistillata* datasets used. The datasets contain preliminary sequence data for the genome, transcriptome, proteome and small RNAome, as of approximately June 2012.

The *Symbiodinium sp.* cultures used were not axenic: thus, care was taken to minimise the amount of bacterial sequences in the resulting datasets. For the genome, contigs that mapped very well to known bacterial sequences were discarded; for the transcriptome and proteome, the RNA extraction step included an additional poly(A)+ selection step, which selectively enriches for eukaryotic mRNA.

Unfortunately, for the small RNAome, there is currently no way of assessing the extent of bacterial contamination due to the dearth of deep sequencing data from other dinoflagellates, but I believe that the contamination should not affect the conclusions drawn from the overall data. From experience, BLAST searches of the small RNA data against the NCBI "nr" database, which contains eukaryotic and bacterial data, tended to match sequences from eukaryotes than prokaryotes. Also, some of the prokaryotic hits were to other known chloroplast or mitochondrial sequences — as expected of a photosynthetic dinoflagellate.

4.2.2 Identification of core proteins required for RNAi

In order to identify homologues of the RNAi machinery in both organisms, sequences from six families of proteins (Argonaute, Dicer, Piwi, Drosha, Pasha and HEN1) were drawn from five model organisms (*H. sapiens*, *D. melanogaster*, *C. elegans*, *S. pombe* and *A. thaliana*) that spanned the major kingdoms of life. These sequences were obtained from UniprotKB (`http://www.uniprot.org`), and clustered into groups with 90% sequence identities to remove near-identical sequences.

These RNAi proteins were then searched against protein sequences from both organisms

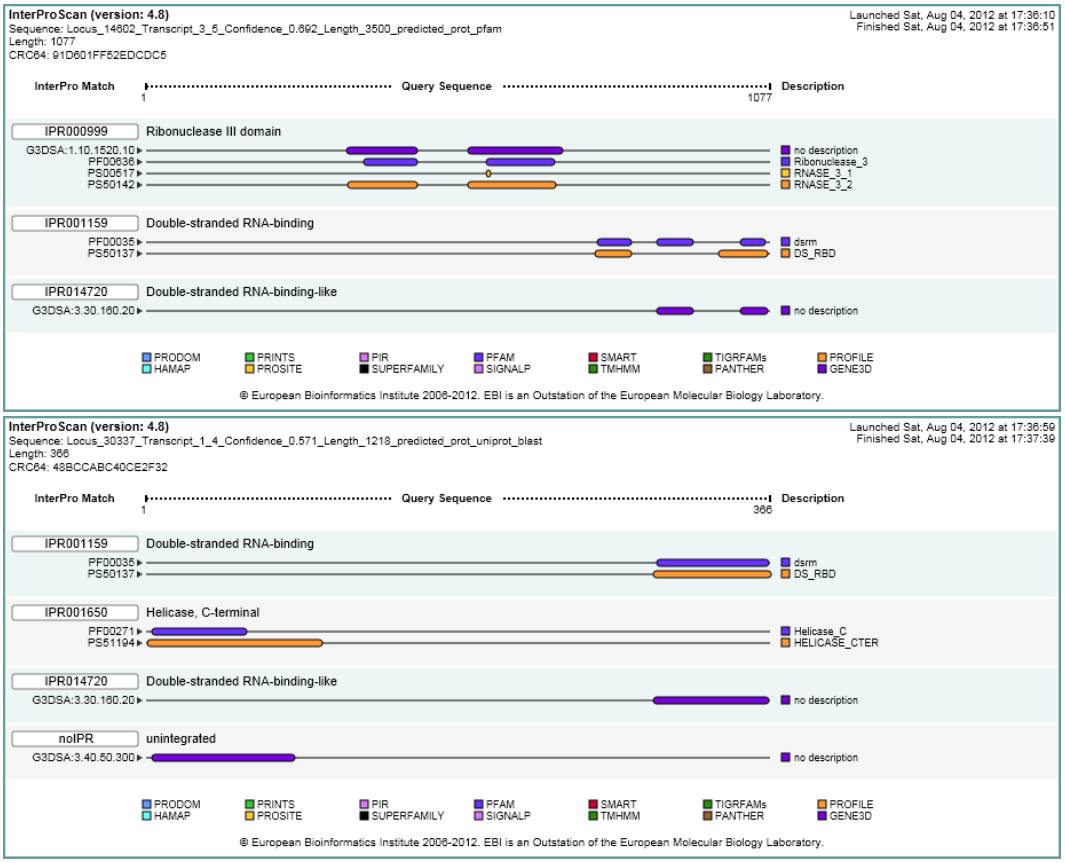


Figure 4.6: Output from InterProScan illustrating the retention of candidate RNAi proteins based on presence of crucial protein domains. Both "Locus_14602" (top) and "Locus_30337" (bottom) are proteins that closely match known Dicer. The pair of RNase III domains crucial for Dicer activity is only present in the former but not the latter. Thus, only the former is still considered as a homologue of Dicer.

using BLASTp. Candidate homologues (BLASTp e-value $<$ 1e-10) of known RNAi proteins were then searched for domains that are required for the catalytic function of the protein using InterProScan (`http://www.ebi.ac.uk/Tools/pfa/iprscan/`) (Zdobnov and Apweiler, 2001; Quevillon *et al.*, 2005; Mulder *et al.*, 2007). These crucial domains are: Paz and Piwi for Argonaute and Piwi proteins; a pair of RNase III domains for Dicer and Drosha; a double-stranded RNA binding domain for Pasha; and a methyltransferase (MTase) domain for HEN1. Candidate homologues are retained based on the presence of the crucial domain — see Fig 4.6 that contrasts a retained homologue of Dicer against a discarded one. Additional support for the inferred function of candidate homologues was obtained by carrying out a reciprocal BLAST search of these candidates against all GenBank protein sequences ("nr").

Using Clustal Omega (Sievers *et al.*, 2011), the candidate homologues were aligned

against known RNAi proteins on a per-family basis. The alignments aided the search for certain strongly-conserved residues in the protein domains associated with RNAi activity. Jalview (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009) was used to view these alignments, while PhyML (Phylogenetic estimation using Maximum Likelihood) (Guindon and Gascuel, 2003; Guindon *et al.*, 2010) was used to infer phylogenetic relationships between aligned proteins. The resulting trees (in Newick format) produced by PhyML were visualised using iTOL (interactive Tree Of Life, `http://itol.embl.de`) (Letunic and Bork, 2007, 2011).

4.2.3 Extraction of small RNA for library generation

The small RNA fractions for *Symbiodinium sp.* and *S. pistillata* were selectively enriched from total RNA extracts using the mirVana kit (Ambion). The extraction of total RNA, and its associated challenges, from *Symbiodinium* has been described in the previous chapter; while the total RNA from *S. pistillata* was kindly provided by our collaborator Didier Zoccola (Centre Scientifique de Monaco).

For *Symbiodinium sp.*, we were also interested in investigating the transcriptional changes that occur during the response to environmental stresses. The list of stresses, full details about the amounts of RNA extracted for each condition, as well as quality control in the form of spectrophotographic measurements and electropherograms, can be found in Table 3.5 in the previous chapter.

4.2.4 Library generation for Illumina sequencing

The creation of all libraries and their subsequent Illumina sequencing were performed by the in-house sequencing facility in KAUST (Saudi Arabia). For all libraries, after the Illumina-provided 5' and 3' adapters were ligated to the small RNAs, a gel was run to size-select molecules that were ∼140–160 bp, which corresponds to small RNA of initial lengths of ∼15–35 bp.

For *S. pistillata*, one small RNA library was created using Illumina's Small RNA Sample

Prep Kit, and sequenced to produce ∼30.5 million reads.

For *Symbiodinium sp.*, two libraries were produced from the small RNA extracts. The first *Symbiodinium* small RNA library was created from pooling roughly equal amounts (∼0.5 µg) of RNA from each of the nine growth conditions using Illumina's Small RNA Sample Prep Kit, and sequenced to produce ∼31.1 million reads. For the second library, the newer TruSeq Small RNA Sample Prep Kit was used, which allowed for multiplexed sequencing — cDNA from each of the nine growth conditions has a 5' primer specific to the condition ligated during library generation. The cDNA was sequenced along four lanes to produce a total of 137 million reads.

Data from the Illumina sequencing were stored as FASTQ files, which contains per-base quality information for the entire length of the read.

4.2.5 Processing of FASTQ reads for analysis

As the raw FASTQ files were extremely large — each file was several gigabytes in size — and contained low-quality reads, the FASTQ reads had to be processed by several scripts to produce a more compact FASTA file that contained high-quality reads for downstream analyses. This pipeline is illustrated in Figure 4.7.

4.2.5.1 Trimming low-quality bases from 3' ends

In the FASTQ file, Phred quality scores are assigned to each base to denote the error rate associated with the sequencing of the base. Mathematically, Phred scores are calculated using the equation:

$$Phred\,score = -10\,log_{10}(Error\,rate)$$

For example, if the error rate associated with the sequencing of a base is 0.001, the Phred score of the base would be 30.

Typically, the base qualities of short reads are higher (Phred score > 30) at the 5' end. As the rate of sequencing errors increases with the length of the sequenced read, the read

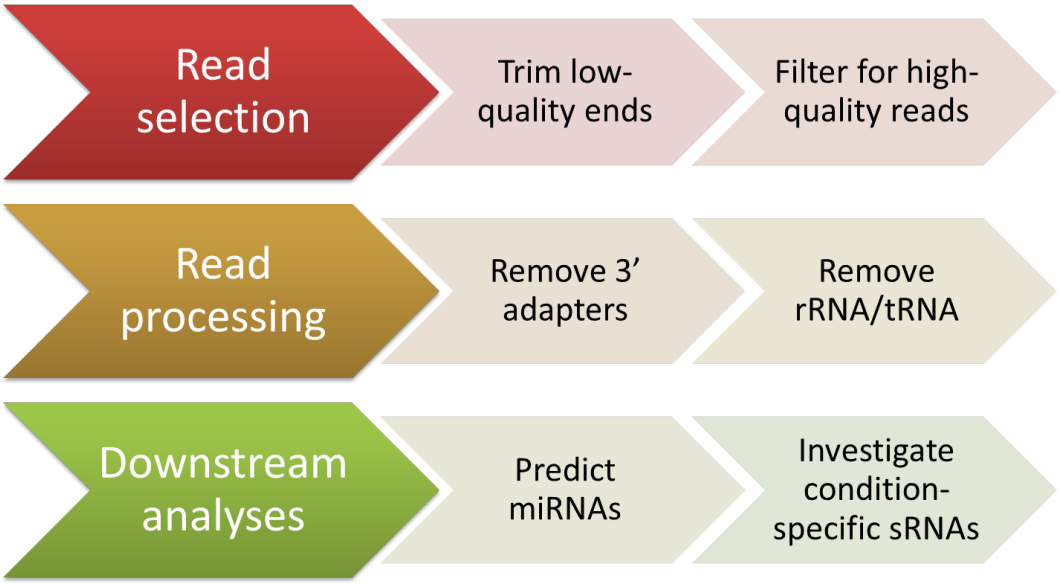


Figure 4.7: Flowchart illustrating the analysis pipeline carried out on raw FASTQ files.

quality of the last few bases can be very poor (Phred < 20). Instead of discarding all reads that contained several low-quality bases, a Perl script written by Joseph Fass (unpublished but in public domain, with slight modifications by me) was used to remove low-quality bases from the end of the read. For a short read that is originally $(m + n)$ nucleotides in length, the script trims off the shortest-possible n nucleotides from the 3' end to produce a read of m nucleotides that satisfies two criteria: the m^{th} nucleotide has a Phred score of > 20, and the average Phred quality across all m nucleotides are > 20. In the example given in Figure 4.8, the last five bases in the short read had a Phred score of 3, while the sixth-last base has a Phred score of 31 (satisfies first criterion). The average Phred score of the resulting trimmed sequence is 26.7 (which satisfies the second criterion).

4.2.5.2 Filtering for high-quality reads

After trimming, a Python script called "filter_fastq_by_overall_quality.py" was written to assess the overall quality of reads. Based on this overall quality, the read was either kept or discarded. As the reads resulting from the trimming steps have varying lengths, the use of average Phred scores as a measure of quality is not very accurate — take for example a 20

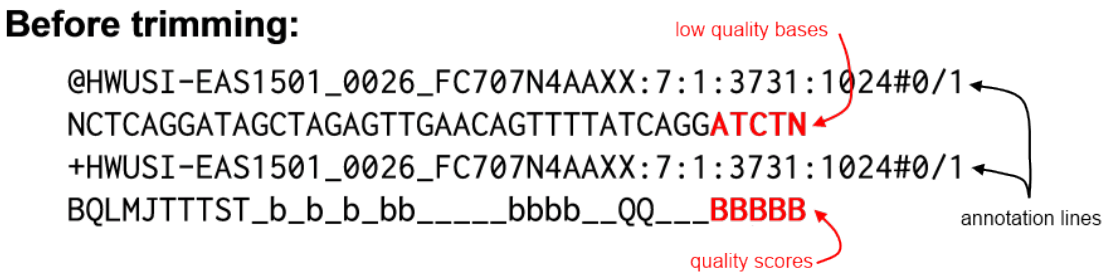


Figure 4.8: Trimming of low-quality bases. In the FASTQ Phred+64 notation, "A" denotes a Phred score of 2, "B" denotes 3, "C" denotes 4 and so on. As there are more quality scores than uppercase alphabets, other symbols are used to denote larger Phred scores. In this figure, "_" denotes a Phred score of 31, while "b", in lowercase, denotes 34. The script trims off bases with low Phred scores (in red), producing a trimmed sequence with an average Phred score of 26.7.

bp long and a 40 bp long read, both only containing bases of Phred score 30 (i.e. each base has a sequencing error rate of 0.001). The probability of a read being completely error-free is higher in the shorter read ($0.999^{20} = 98.0\%$) than that in the longer one ($0.999^{40} = 96.1\%$).

Thus, for each read, the probability of it being completely error-free was calculated directly from the qualities of each base. Reads that had a $< 98\%$ chance of being error-free were discarded, the rest being retained in a FASTQ file that only contained high quality reads (see Figure 4.9). Across all three libraries, on average, $\sim$20% of the reads are discarded due to the quality thresholding.

4.2.5.3 Removing 3' adapters from reads

Due to the short length of the small RNA reads, the resulting sequenced read is often longer than the RNA itself. Thus, portions of the 3' adapter are present in the resulting read. A program called Cutadapt v1.0 (Martin, 2011) was used to remove the leading bits of the 3' adapter (and, occasionally, trailing bits of the 5' adapter) from the short reads. The choice

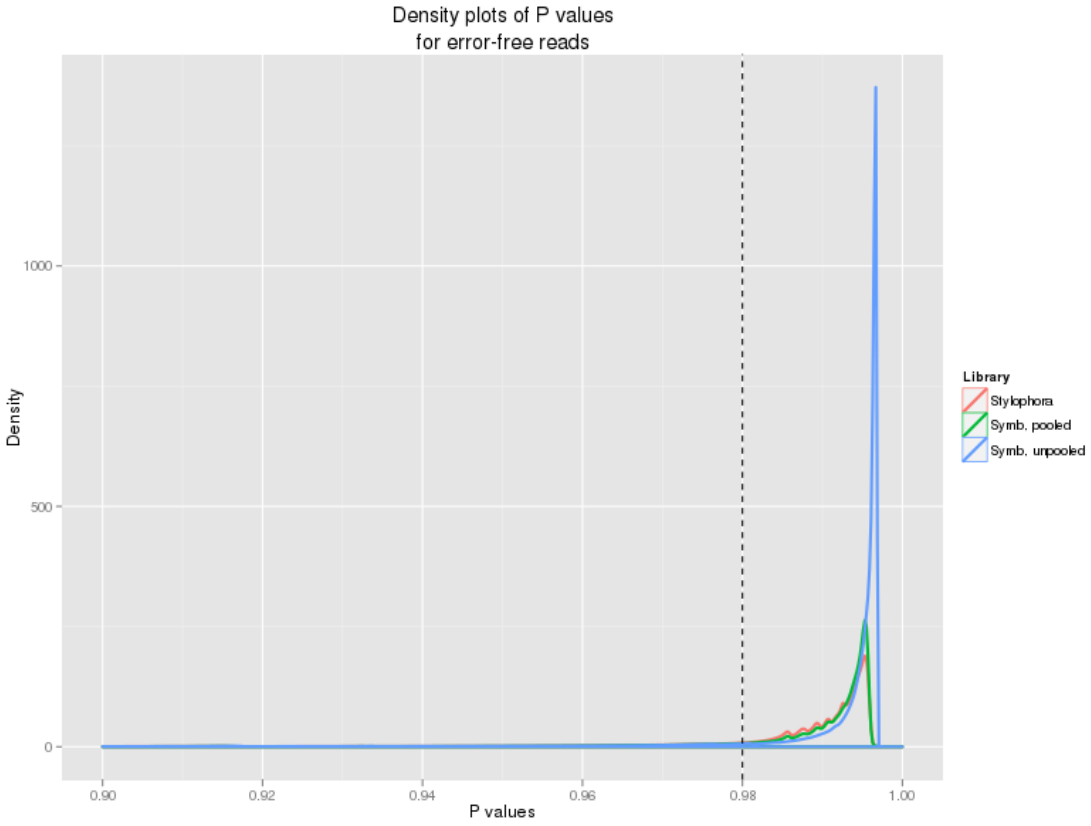


Figure 4.9: Density plots of short reads being error-free for all three small RNA libraries. The vertical line shows the quality threshold where reads were discarded if P =< 0.98, and kept if P > 0.98.

of non-default parameters for Cutadapt is listed in the Appendix (Section 5.3.1).

4.2.5.4 Filtering rRNA-, tRNA- and mRNA-related short reads

Although the small RNA libraries were created from RNA extracts with the small RNA (< 200 bp) fraction selectively enriched, fragments of rRNA and tRNA sequences were still observed in the short read data. Short reads containing portions of mRNA sequences were also found in the short read data, albeit at much lower frequency than rRNA- and tRNA-related reads.

In order to remove these short reads from the downstream analyses, Velvet (Zerbino and Birney, 2008) was used to assemble the short reads into contigs (at hash length of k=25). A general search of these contigs against "nt" (nucleotide collection in NCBI) and the assembled transcript sequences ("datafreeze" version for *Symbiodinium sp.*; "beta_anno" version for *S. pistillata*) was carried out to find out which contigs had short reads that matched known rRNA, tRNA and transcripts sequences. These short reads were then filtered out of the short read dataset. Table 4.3 and 4.4 lists the pre- and post-filtering of read counts across the three available small RNA libraries.

	***Symbiodinium sp.* unpooled library**	***Symbiodinium sp.* pooled library**	***S. pistillata***
Pre-filtering	103368244	25422337	23830932
Post-filtering	99624754	23227235	21625195
% filtered	3.62%	8.63%	9.26%

Table 4.3: Details for the reads filtered out from the three overall libraries. Despite the larger overall size of the unpooled *Symbiodinium sp.* library, it had a lower percentage of reads which are fragments of rRNA, tRNA or transcripts.

	16C	**20g**	**36C**	**4C**	**60g**	**DC**	**DS**	**HS**	**noon**
Pre-filtering	12593192	15378342	10658877	17193220	8725667	10662995	6131326	8433048	13591577
Post-filtering	12144083	14759514	10220289	16413772	8432981	10287027	6030793	8263901	13072394
% filtered	3.57%	4.02%	4.11%	4.53%	3.35%	3.53%	1.64%	2.01%	3.82%

Table 4.4: Breakdown of *Symbiodinium sp.* unpooled library into its constituent condition-specific reads. The number of reads (∼2–4%) that were fragments of rRNA, tRNA or transcripts were fairly consistent across the nine different conditions.

4.2.6 miRNA prediction: miRDeep2 as program of choice

With the advent of high-throughput sequencing, miRNA detection is now possible at unprecedented sensitivities over traditional gel-based methods. However, the task of accurately identifying *bona fide* miRNAs from short read data remain challenging (Friedländer *et al.*, 2012). Recent genome papers have started to conduct *in silico* genome-wide predictions of miRNAs due to the increased recognition of the role of sRNAs in regulating gene expression, but the methods employed for the predictions are not standardised (e.g. for *N. vectensis*, Grimson *et al.* (2008) wrote an in-house prediction pipeline; for the butterfly *Heliconius melpomene*, Dasmahapatra *et al.* (2012) used miRCat (Moxon *et al.*, 2008)).

Currently, there are many software packages that handle genome-wide miRNA predictions: DSAP (Huang *et al.*, 2010a), miRanalyzer (Hackenberg *et al.*, 2009), miRCat (Moxon *et al.*, 2008), miRDeep v1 (Friedländer *et al.*, 2008) and v2 (Friedländer *et al.*, 2012), miRDeep-P (Yang and Li, 2011) and many others. These programs differ in several key aspects, namely target genome (restricted to a pool of model organisms, or flexible), accessibility (web-based or downloadable binaries) and prediction algorithms.

While most of the prediction pipelines are untested against each other, Williamson *et al.* (2012) has assessed the accuracy and performance of four programs — DSAP, miRanalyzer and both version of miRDeeps. Among all four programs, miRDeep v1 and v2 were shown to be more sensitive at predicting miRNAs. Although miRDeep v1 was slightly more sensitive than v2 in the analysis carried out by Williamson *et al.* (2012), miRDeep v2 was chosen because of the significant performance improvements (it is $\sim$4 times faster than its predecessor) and the reported improvements on the predecessor's algorithm (Friedländer *et al.*, 2012). Also, the binaries for miRDeep (-P and v2) can be compiled to run locally — considering the multi-gigabyte filesizes involved, locally run programs are faster and more reliable compared to remotely-hosted ones.

In the genome-wide miRNA predictions for *Symbiodinium* and *S. pistillata*, miRDeep v2 and miRDeep-P were used (detailed explanations are in the species-specific subsections below). The basic principle behind both programs is the same — the short reads are first

mapped to the genome, and regions that contain an abundance of reads are then folded to ascertain whether they resemble the hairpin structure for miRNA precursors (pre-miRNAs).

However, miRDeep-P relaxes several hardcoded constraints present in miRDeep2 as plant miRNA biogenesis differs from its animal counterpart in several respects. Plant pre-miRNAs are much longer and of more variable length; plant miRNAs tend to belong to larger paralogous families, with members coding identical or near-identical miRNAs. For the former, miRDeep-P relaxes the length restriction of potential precursor sequences, which allows for the detection of potential miRNAs in longer hairpin sequences; for the latter, miRDeep-P allows for the miRNAs to be detected up to 15 times in the genome, three times the original restriction (which was put in place to significantly reduce false positives) (Yang and Li, 2011; Friedländer *et al.*, 2012).

4.2.6.1 *Symbiodinium sp.*

As the molecular machinery of miRNA biogenesis remain uncertain in dinoflagellates, it remains unknown whether miRNA precursors in dinoflagellates resemble that in plants or animals. Two miRNA prediction programs were used in predicting miRNAs: miRDeep-P (Yang and Li, 2011) and miRDeep2 (Friedländer *et al.*, 2012), the former being more suited for detecting plant miRNAs, and the latter animal miRNAs.

The number of mature miRNAs predicted by both programs across both libraries is shown in Table 4.5.

	Pooled library	**Unpooled library**	**Overlap**
miRDeep2	70	86	31
miRDeep-P	55	65	18
Overlap	41	42	15

Table 4.5: Summary of mature miRNA prediction using miRDeep2 and miRDeep-P for both *Symbiodinium* libraries.

We decided to base our downstream analyses on the common set of 31 mature miRNAs across both libraries predicted from miRDeep2 only, as miRDeep2 brings about many prediction improvements over miRDeep version 1, which served as the code base for miRDeep-P.

The improved miRDeep2 algorithm has an experimentally-verified ∼99% accuracy in identifying known miRNAs across seven animal clades (Friedländer *et al.*, 2012). This high accuracy makes miRDeep2 stand out amongst the numerous miRNA prediction programs — an independent assessment of miRDeep2 against other programs, such as miRanalyzer and DSAP, showed that miRDeep2 was more accurate at predicting *bona fide* miRNAs (Williamson *et al.*, 2012). The large overlap (∼70% of predictions) of predicted mature miRNAs from both programs shows that miRDeep2 is able to reproduce most of the predictions from miRDeep-P despite the precursor length restrictions, with the advantage of producing ∼30% more predictions than miRDeep-P.

As lowly-expressed reads can be quite variable between individual runs, the miRNAs predicted from both libraries represent a set of miRNA predictions that has a higher chance of being *bona fide* miRNAs in *Symbiodinium*. Accuracy is valued higher than sensitivity at this exploratory stage.

4.2.6.2 *Stylophora pistillata*

As *S. pistillata* is a metazoan, only miRDeep2 was used for the genome-wide prediction of this coral species. The prediction pipeline was ran on our sole small RNA library.

4.2.7 Identification of condition-specific short reads in *Symbiodinium sp.*

In order to reduce the rate of false positives, we decided to focus on abundant short reads that exhibited a condition-specific expression pattern. Several precautions were taken to ensure a fair estimate of read abundance: firstly, read counts from the nine different conditions were normalised against each other to allow for comparisons of specific reads across conditions; secondly, while coming up with a reasonable and suitable cut-off for abundance, we discovered that there were reads which were fairly similar to other reads in the dataset (e.g. single- to several-nucleotide offsets, or single-nucleotide substitutions). Clustering these similar sequences together would thus give a fairer estimate of the abundance of these sequences.

4.2.7.1 BaySeq normalisation of reads across conditions

The choice of normalisation procedure has been shown to have a great effect on the capability of detecting differential expression across multiple datasets. Traditionally, most normalisation procedures involve the global scaling of individual read counts by the number of total reads in that dataset. However, this scaling is heavily skewed by the presence of a small proportion of highly-expressed reads (e.g. in Bullard *et al.* (2010), 50% of the total read counts in their datasets can be attributed to 5% of the genes). Direct scaling based on total read counts makes the inherent assumption that abundant reads have similar expression levels across different conditions, which is, biologically speaking, not guaranteed.

In Bullard *et al.* (2010), the authors proposed a quantile-based scaling method in normalising reads — instead of scaling by the sum of all reads in a lane, read counts are instead scaled against the value of the upper quartile (75th percentile). It has been shown that quantile normalisation improves the sensitivity at detecting differentially expressed genes.

This quantile-based scaling method has been implemented in BaySeq (Hardcastle and Kelly, 2010), thus the program was used to normalise read counts across conditions.

4.2.7.2 Clustering of similar reads via cd-hit-est

Post-normalisation, reads with similar sequences were clustered using cd-hit-est (Li and Godzik, 2006) (see Appendix, Section 5.3.1 for the list of non-standard parameters used for this step). Briefly, while creating a new cluster, the algorithm selects for the longest sequence as the "representative sequence", and compares other reads in the dataset to this sequence. If the similarity is above a predefined threshold, reads are clustered together. A new representative sequence is chosen out of the unclustered reads, and the clustering procedure begins anew.

4.2.7.3 Identifying condition-specific abundant short reads

After clustering, we selected for clusters that had more than 1,000 reads combined across all nine conditions, resulting in 1,772 clusters that were above this abundance threshold.

In order to analyse the patterns of short read expression, the hierarchical clustering algorithm in MeV v4.8 (MultiExperiment Viewer, `http://www.tm4.org/mev/`) (Saeed *et al.*, 2003, 2006) was used to visualise and group short reads with similar patterns of expression. For each read, the condition-specific read counts were converted into Z scores (i.e. a value of "2" means that the read count from the specific condition was 2 standard deviations away from the mean) to produce an expression pattern independent of the overall total read count.

4.3 Results and discussion

4.3.1 Identification of core proteins required for RNAi

In *Symbiodinium sp.*, three candidate Argonautes, one Dicer and one HEN1 were discovered; in *S. pistillata*, three candidate Argonautes, eight Dicers, one Piwi, one Pasha and two HEN1 were present. Tables 4.6 and 4.7 summarises the key metrics (matches to known RNAi families, presence of protein domains crucial for catalytic activity, and the reciprocal BLAST search against all annotated proteins) for the candidate RNAi proteins. Full sequences of these proteins can be found in the Appendix (Section 5.3.2).

The per-family alignments of candidate homologues against known sequences reveal the striking conservation of functionally important amino acid residues located within key protein domains (PAZ and Piwi domains in Argonaute and Piwi; both RNaseIII domains in Dicer; the dsRNA-binding domain in Pasha; and the methyltransferase domain in HEN1). The exact identities of these conserved residues will be elaborated in the next few subsections. The strong conservation of key protein domains suggests the presence of a functional RNAi machinery in both organisms.

4.3.1.1 Argonaute/Piwi family

For Argonaute and Piwi proteins, two key protein domains are conserved — the PAZ and Piwi domains. Although the function of PAZ remains unclear, Lingel *et al.* (2003) carried out a mutational analysis of several highly-conserved residues in the PAZ domain implicated

Protein annotation (shortened)	Length	Hits (E < 1e-10) to known RNAi families						% identity (average)	Key protein domains present					Inferred function	Reciprocal BLAST to "nr": Top hit	
		Ago	Dicer	Drosha	Pasha	Piwi	HEN1		Paz	Piwi	RNaseIII	dsRBD	MTase		Annotated function	Organism
Locus_14602	1077		2					24.9			2	2		Dicer	Putative Dicer-like	Fungus
Locus_1844	823	19				6		24.3	1	1				Argonaute	Piwi-like protein 1	Platypus
Locus_19351	1337						4	29.7					1	HEN1	Methyltransferase	Streptomycete
maker-2878763	682	20				6		26.3	1	1				Argonaute	Argonaute-2-like	Demosponge
maker-3727646	796	16				6		24.5	1	1				Argonaute	Seawi (sea urchin piwi)	Sea urchin

Table 4.6: RNAi-associated candidate proteins in *Symbiodinium sp.* For the reciprocal BLAST search, hits to vague annotations such as "predicted protein" were removed in favour of hits that contain (in some cases, predicted) protein function. Although "Locus_1844" and "maker-3727646" better match known Piwi sequences than Argonaute sequences, the absence of an enriched 25–30 nt fraction starting with 5'-U indicates that piRNAs are not present in *Symbiodinium sp.* (data not shown).

Protein annotation (shortened)	Length	Hits (E < 1e-10) to known RNAi families						% identity (average)	Key protein domains present					Inferred function	Reciprocal BLAST to "nr": Top hit	
		Ago	Dicer	Drosha	Pasha	Piwi	HEN1		Paz	Piwi	RNaseIII	dsRBD	MTase		Annotated function	Organism
maker-6213567	990		9					30.4	1		2			Dicer	Dicer	Zebra finch
maker-6243026	792	20				6		35.4	1	1				Argonaute	Argonaute-2	Human
maker-6462987	706		9					31.5	1		2			Dicer	Dicer-1-like	Pea aphid
maker-6574134	1021		11					31.9	1		2	1		Dicer	Dicer-like	Demosponge
maker-6582237	1008		9					30.2	1		2			Dicer	Dicer	Zebra finch
maker-6624208	508						4	32.2					1	HEN1	piRNA MTase-like	Acorn worm
maker-6730085	455	1				6		38.0	1	1				Piwi	Piwi	Sponge
maker-6734235	775	19				2		32.2	1	1				Argonaute	Argonaute 1	Sea urchin
maker-6743944	524		9					33.6	1		2	1		Dicer	Dicer-1	Sea anemone
maker-6767481	661		9					36.9			1	1		Dicer	Dicer-1	Sea anemone
maker-6776023	1378		19					39.9	1		2	1		Dicer	Dicer-like	Demosponge
maker-6778374	840	20				6		35.6	1	1				Argonaute	Argonaute-2-like	Human
maker-6785799	550						4	32.2					1	HEN1	piRNA MTase-like	Acorn worm
maker-6789121	529				3			36.7				1		Pasha	DGCR8 (Pasha)	Wild boar
snap_masked-6652293	604		9					31.7	1		2			Dicer	Dicer-1	Sea anemone

Table 4.7: RNAi-associated candidate proteins in *S. pistillata*. For all candidates, the top hit from reciprocal searches (again, excluding vague descriptors like "predicted protein") lends strong support to the inferred functions of those proteins. All of the top hits have metazoan origins as well.

in RNA binding. Figure 4.10 shows that several of these key residues (red asterisks in the figure) are conserved in both *Symbiodinium sp.* and *S. pistillata*.

The Piwi domain in Argonaute contains a strongly-conserved DDE motif, which is present in the active site and contributes to the slicing activity of the ribonuclease (Song *et al.*, 2004). This motif is conserved across most of the candidate homologues, as shown in Figure 4.11.

Using PhyML, a phylogenetic tree was constructed from the aligned Argonaute sequences (shown in Figure 4.12). As expected, candidate homologues from *Symbiodinium sp.* and *S. pistillata* generally showed close matches to homologues that originated from the same species. Interestingly, two candidate homologues from *Symbiodinium sp.* ("Locus_1844_SYMB", "maker-3727646_SYMB") and one from *S. pistillata* ("maker-6730085_STYPI") showed closer matches to known Piwis than Argonautes. However, the presence of Piwi in *Symbiodinium sp.* is unlikely due to the absence of an enriched 25–30 nt fraction with uridine as the first base; for *S. pistillata*, this enrichment is present, and the candidate homologue had considerably more matches to known Piwi sequences than Argonaute ones, making it the only candidate likely to be a Piwi protein.

4.3.1.2 Dicer proteins

The "dicing" activity of Dicer, which generates mature miRNA from its double-stranded precursor, depends on a pair of RNase III domains located near the C-terminus of the protein. Each of the two domains contain key acidic residues that coordinates with a divalent Mg^{2+} ion, which is essential for the activity of the ribonuclease (Lee *et al.*, 2004b). The conservation of these acidic residues in the first and second RNase III domains of the candidate Dicers are shown separately in Figures 4.13 and 4.14.

A phylogenetic tree was also constructed for the aligned Dicer sequences using PhyML (see Figure 4.15). As seen in the tree, the candidate Dicer homologues from *S. pistillata* clustered best with sequences from the same organism and were closely related to other metazoan sequences; on the other hand, the sole Dicer from *Symbiodinium sp.* shows greater similarity to bacterial Dicer sequences than plant or metazoan sequences.

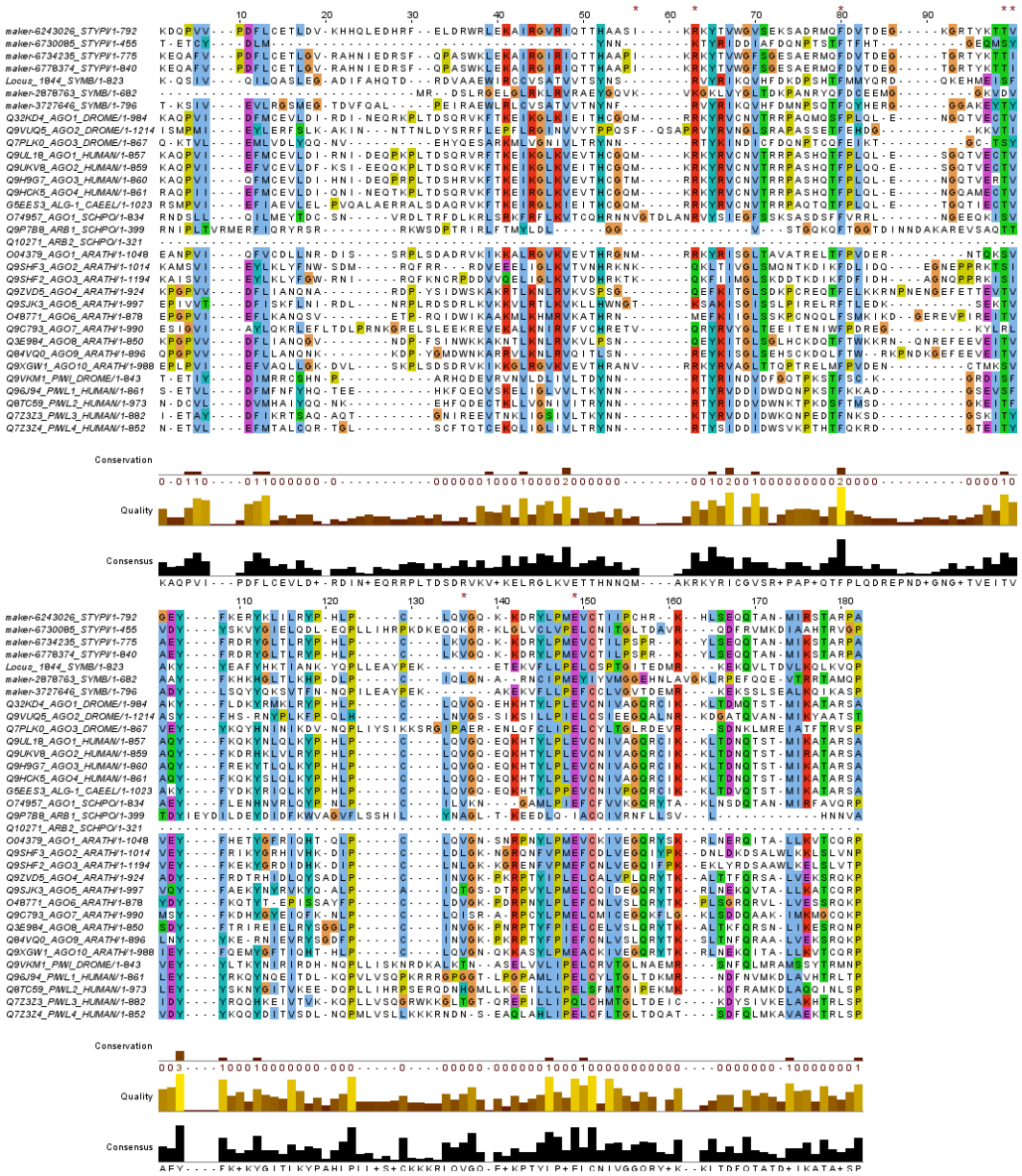


Figure 4.10: Graphical alignment of the PAZ domains in Argonaute and Piwi proteins. Of note are the strong conservation of glutamate (E) at position 149 (mutants produce insoluble protein) and phenylalanine (F) at position 80 (required for RNA binding). However, the phenylalanine at position 56 in *D. melanogaster* AGO2 (also required for RNA binding) was not conserved at all. Key residue positions were obtained from Lingel *et al.* (2003).

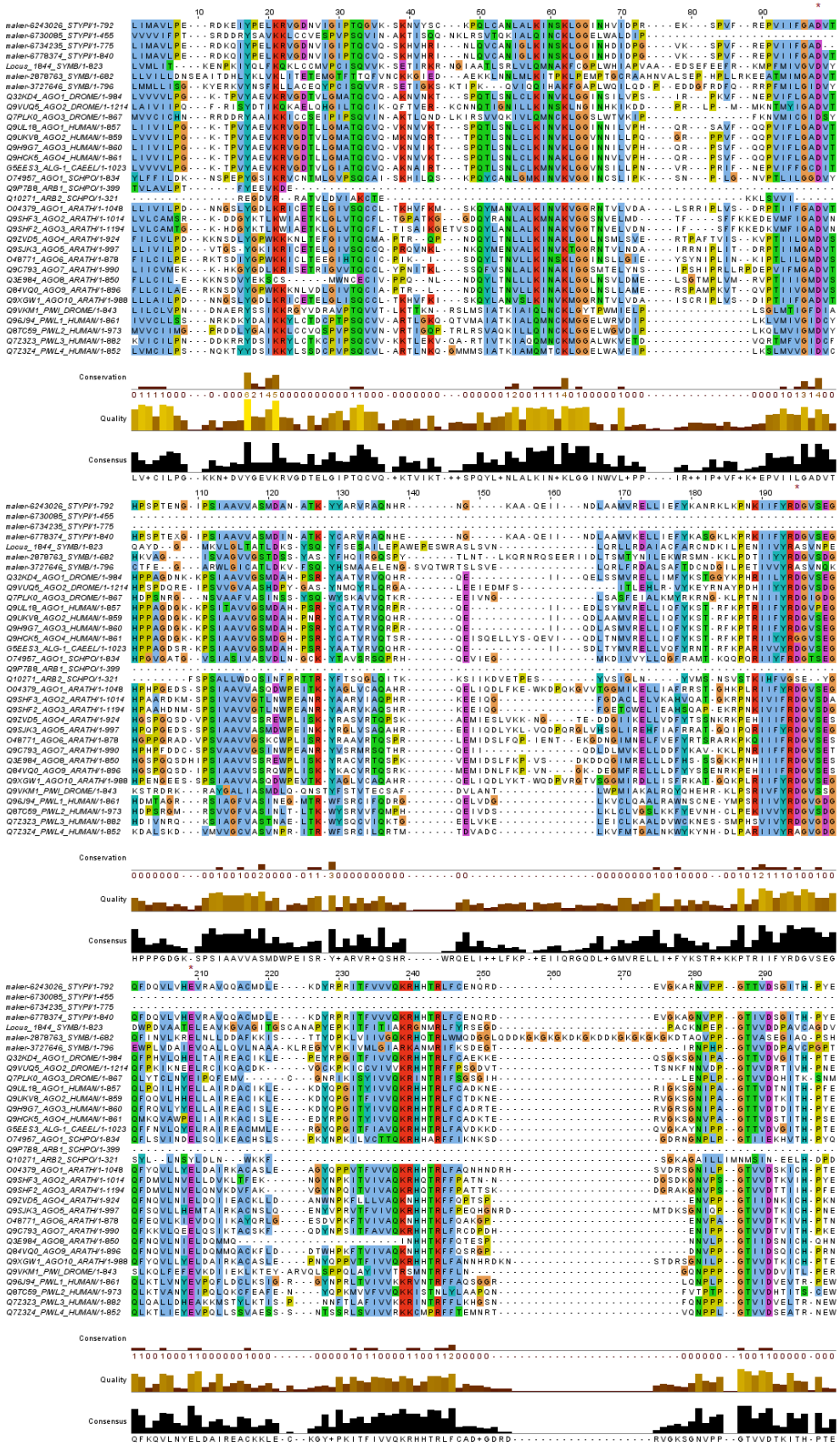


Figure 4.11: Graphical alignment of the Piwi domains in Argonaute and Piwi proteins. The DDE motif is absent in two *S. pistillata* candidate, most likely due to the protein sequences being incomplete (the protein annotations are still work-in-progress).

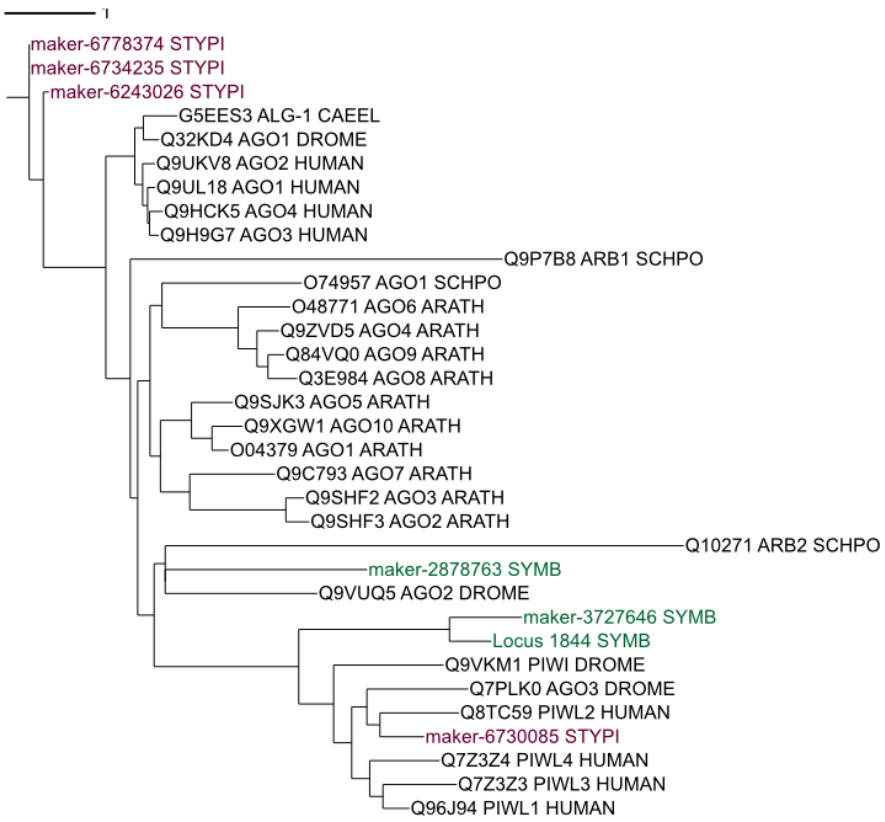


Figure 4.12: Phylogenetic tree constructed for Argonaute/Piwi proteins produced by PhyML. *Symbiodinium sp.* candidate homologues are in green, while *S. pistillata* ones are in red. The clustering of Piwi proteins with other Piwis, and Argonautes with other Argonautes has been observed previously (e.g. in Carmell *et al.* (2002); Murphy *et al.* (2008)). "ARATH": *A. thaliana*; "CAEEL": *C. elegans*; "SCHPO": *S. pombe*; "DROME": *D. melanogaster*.

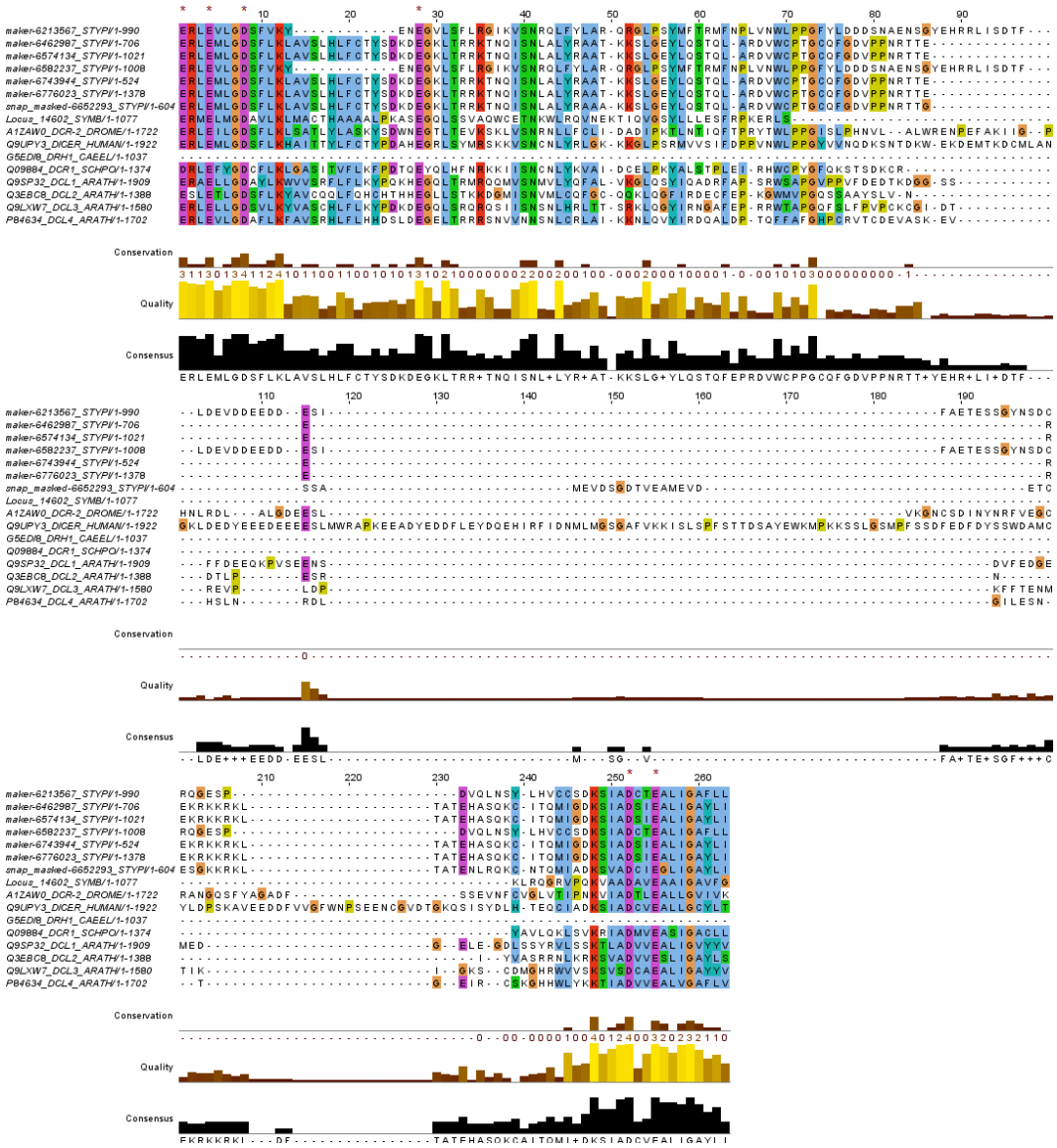


Figure 4.13: Graphical alignment of the first RNase III domain in Dicer proteins. Remarkably, nearly all of the aspartate (D) and glutamate (E) residues involved in the coordination of a divalent metal cation show high levels of conservation across the candidate homologues and known sequences.

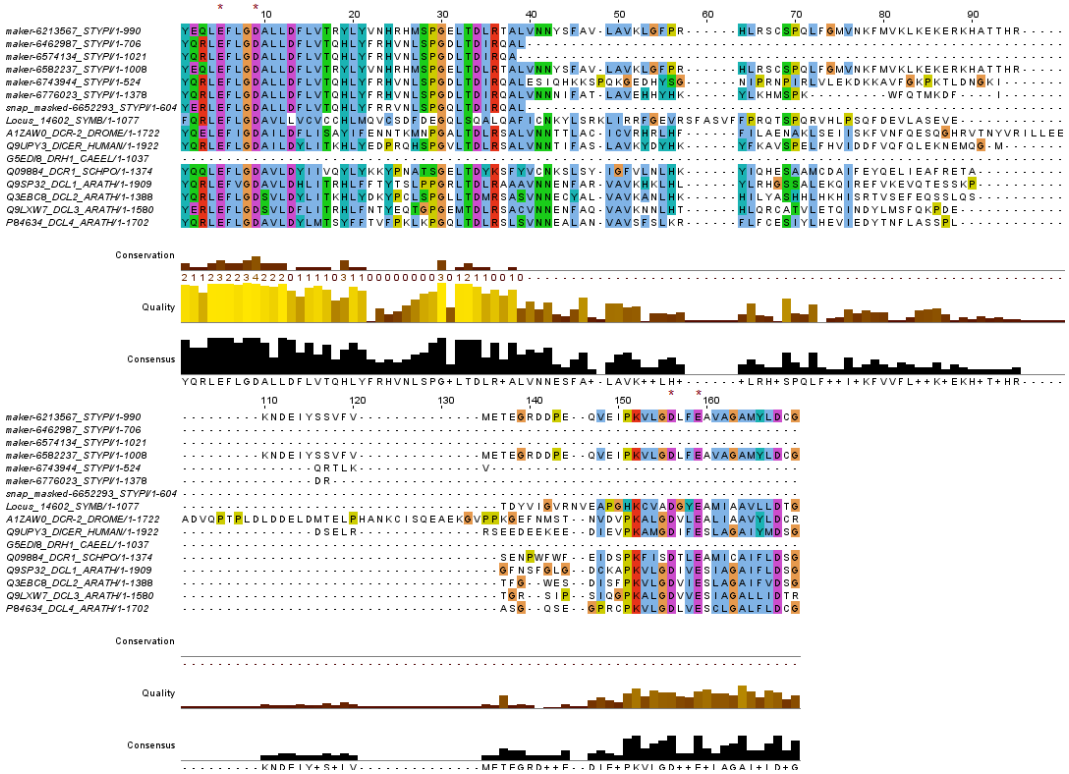


Figure 4.14: Graphical alignment of the second RNase III domain in Dicer proteins. Most of the aspartate (D) and glutamate (E) residues involved in the coordination of a divalent metal cation are conserved. Four *S. pistillata* candidates contain truncated RNase III domains that, despite the truncation, contain the first two key residues and align well to known sequences.

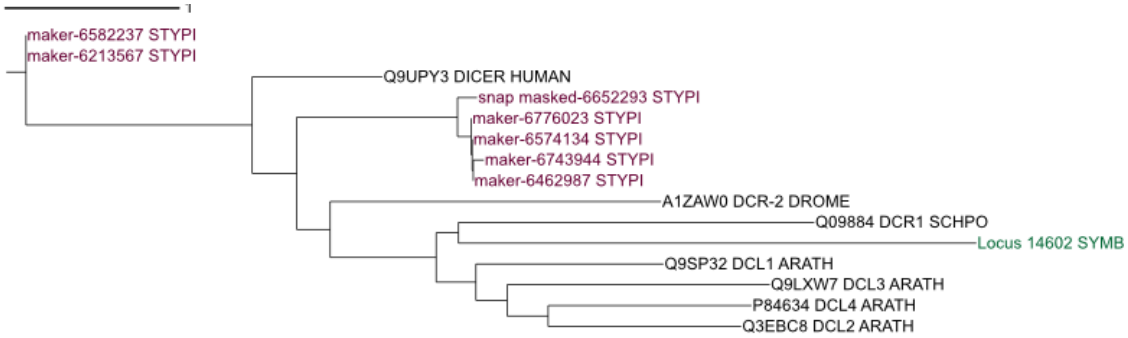


Figure 4.15: Phylogenetic tree constructed for Dicer proteins produced by PhyML. *Symbiodinium sp.* candidate homologues are in green, while *S. pistillata* ones are in red. "ARATH": *A. thaliana*; "SCHPO": *S. pombe*; "DROME": *D. melanogaster*.

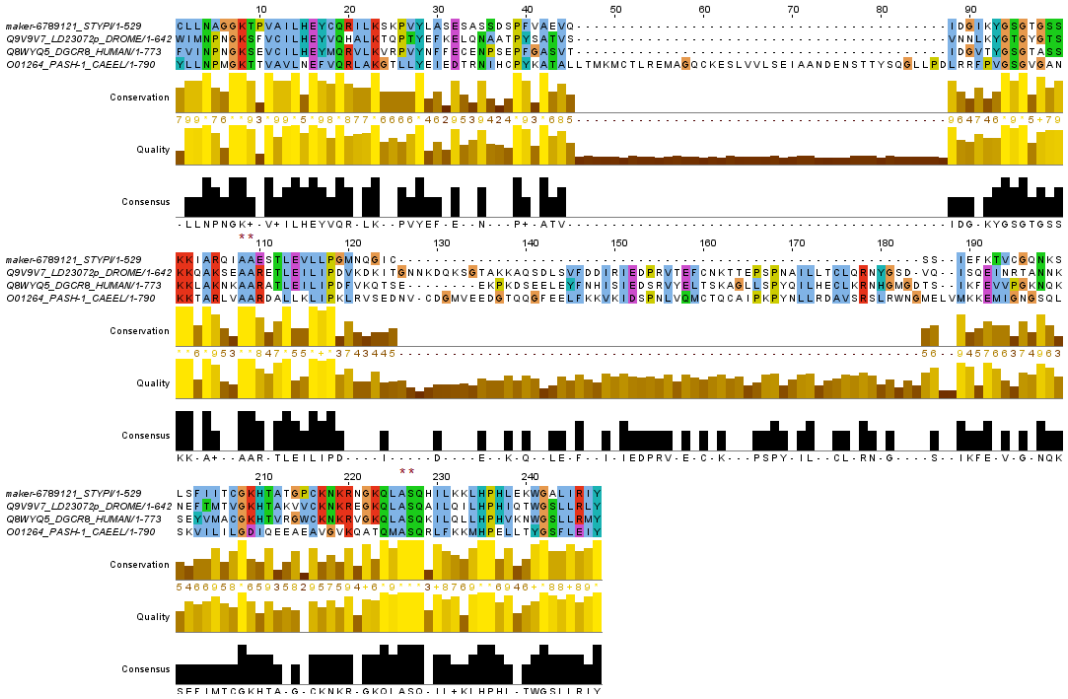


Figure 4.16: Graphical alignment of the dsRNA-binding domain in Pasha. The key alanine/alanine pair (positions 108 and 109) and alanine/serine pair (positions 226 and 227) is present in *S. pistillata* candidate Pasha.

4.3.1.3 Pasha

Pasha, known also as DGCR8 in vertebrates, is an essential cofactor for Drosha. Collectively, both proteins are involved in the generation of pre-miRNA from pri-miRNA in the nucleus. Pasha, with its dsRNA-binding domain, associates and stabilises the pri-miRNA for the endonucleolytic activity of Drosha (Yeom *et al.*, 2006). The binding of dsRNA is dependent on two regions of conserved residues (two contiguous alanines in the first region, and a contiguous alanine/serine pair in the second region), both of which are present in the candidate Pasha from *S. pistillata* (see Figure 4.16). We have yet to find any Drosha that associates with this Pasha, but we remain hopeful that future improvements in the *S. pistillata* protein annotations will reveal the presence of Drosha.

4.3.1.4 HEN1

HEN1 (HUA ENHANCER 1) is a methyltransferase protein involved in the maturation process of some sRNAs, such as siRNAs and miRNAs in plants; piRNAs in metazoans and siR-

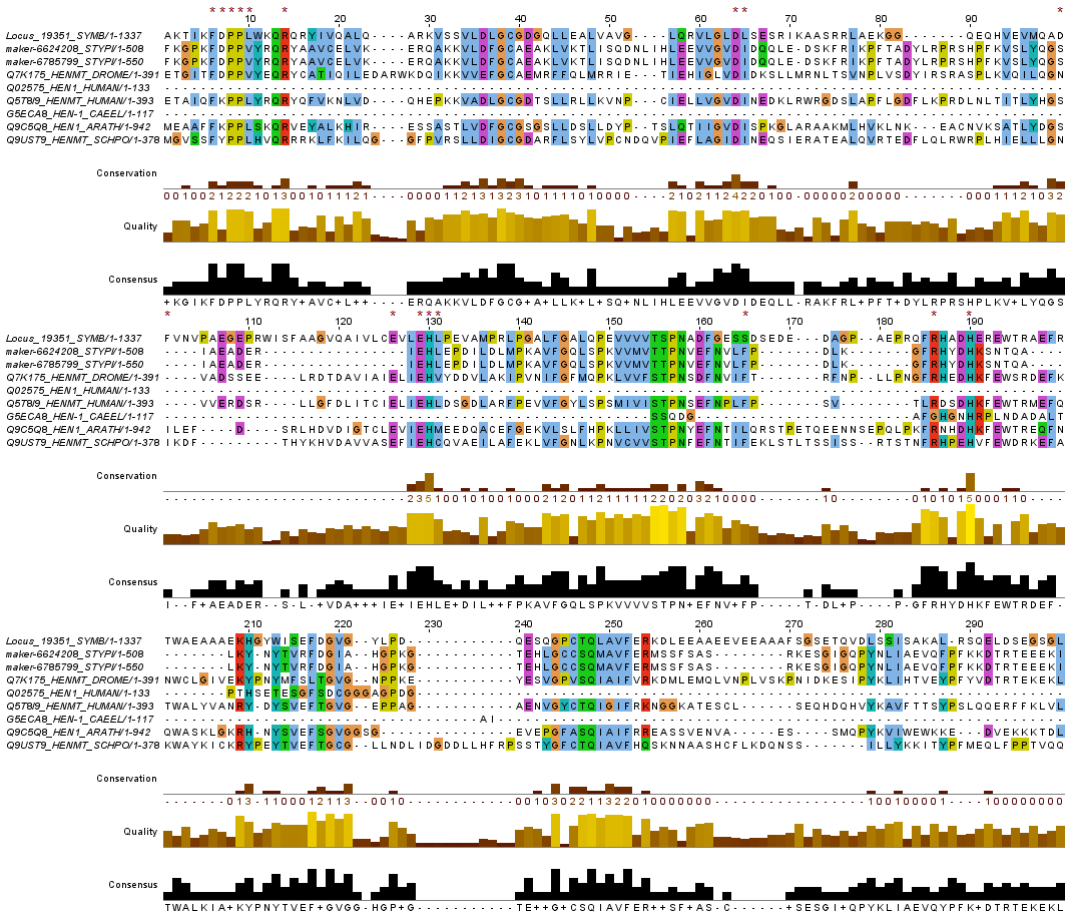


Figure 4.17: Graphical alignment of the methyltransferase domain in HEN1. The residues involved in Mg^{2+} coordination (positions 126, 129, 130 and 190) are well-conserved across the aligned sequences; residues associated with the cofactor AdoHcy and 3' terminus are less well conserved (other positions marked by a red asterisk).

NAs in *Drosophila*, by catalysing the 2'-O-methylation of the 3' terminal nucleotide (Huang *et al.*, 2009). In Arabidopsis, the methylation of miRNA and siRNA ends prevents the uridylation of the 3' terminus, a process used to mark RNA for degradation (Li *et al.*, 2005).

Crystallographic study of the methyltransferase domain in HEN1 has identified several residues that recognise the 3' terminus of sRNAs, coordinate with Mg^{2+} ions or associate with the cofactor adenosyl-L-homocysteine (AdoHcy) (Huang *et al.*, 2009). Many of these residues are conserved in *Symbiodinium sp.* and *S. pistillata* candidate HEN1 homologues as well (see Figure 4.17).

4.3.2 Genome-wide miRNA prediction

Based on the high likelihood that a functional RNAi machinery is present in both organisms, miRNAs have been predicted for both species using miRDeep2. These predicted miRNAs are listed in Tables 4.8 and 4.9.

4.3.2.1 *Symbiodinium sp.*

None of the predicted *Symbiodinium* miRNAs had homologues in other species when compared against all mature miRNA sequences in miRBase (release 18), likely due to the absence of miRNA data from other dinoflagellates. 21 distinct miRNAs were identified from both of the *Symbiodinium sp.* libraries used in the prediction.

4.3.2.2 *Stylophora pistillata*

While the vast majority among the 46 predicted miRNAs were novel, two of the miRNAs predicted in *S. pistillata* matched known miRNAs. The miRNA with the temporary annotation of "6773118" is an exact match of nve-miR-2023 (*N. vectensis* miR-2023), while "6770795", the predicted miRNA with the highest miRDeep2 score, is very similar to known miR-100 family of sequences. For clarity, these two candidate miRNAs in *S. pistillata* will be referred to as spi-miR-2023 and spi-miR-100 respectively, in accordance with miRNA naming conventions. Clustal Omega alignments of these two candidate miRNAs against known sequences are shown in Figures 4.18 and 4.19 respectively.

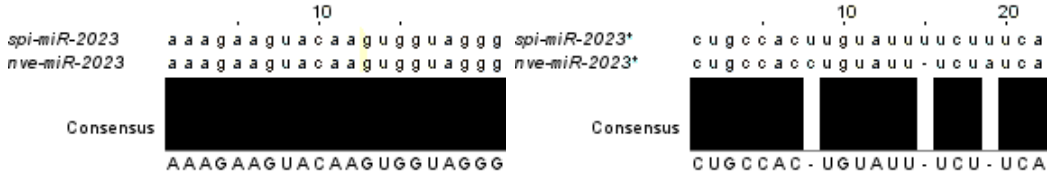
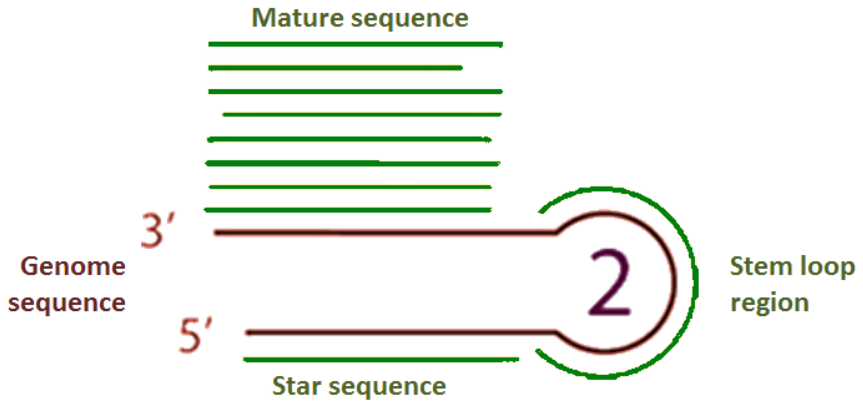


Figure 4.18: Alignment of spi-miR-2023 against nve-miR-2023. The mature sequence is shown on the left, while the star sequence is on the right.

Although nve-miR-100 has been identified in two separate studies, both of which utilised next-generation sequencing of short reads to identify miRNAs in basal metazoans, Wheeler



Genome annotation	Mature miRNA sequence	Star miRNA sequence	miRDeep2 score	
			Unpooled	Pooled
3809624	cgggacucgauucggagggugc	acccuccgaaucgagucucuga	2161.3	1380.8
3809352	ucgaacuuucaggaauaguauc	uaccgucccugagaguucaaug	1211.5	2818.5
3734769	ucuuuaagauauagcucgcggc	cgcaagccauaucuucaagacc	842.7	2648.5
3727297	ugauguacaucgauugaucgac	cgaucaaucgaugugcaucauu	540.1	687
3792382	aguauauuugcugauccgucag	ugacggaucaucaaaucuacauc	485.8	3466.2
3687860	caacgagauuggccuucugugc	acagaaggccaaucucguugcu	458.1	3072
3687860	caacgagauuggccuucugugc	acagaaggccaaucucguugcu	458.1	3072
3690600	acuuagaacucuccuacgaggg	cucguaggagaguucuaagucg	400.7	435.3
3808648	acagggaccugaacagaaaaac	uuuucuguucaagucacgguag	316.8	1291.6
3762090	uuugucgcuccaaccaucacga	gugaugguuggggcagcaaaga	225.4	1362.3
3804066	uccgcaugaagggaaggcuugg	aaguccugcccuucauguguaaa	148	445.3
3791867	uagauggcaagcuggaagaaga	uuguuccagcucgacaucuacc	139.9	309.7
701826	gcgcgguugcugccaucuguugc	acagauugcggcaagcgugcag	133.3	196
701825	gcgcgguugcugccaucuguugc	acagauugcggcaagcgugcag	132.3	195.5
3711874	gaggaugcugaucauucacugg	aguaaaugaucagcauccucca	102.8	103.3
3791567	guuugacucgugccuucccccg	ggggaaggcaucagucgaaucu	74.1	809.3
3801694	ucaagaauugaggaugccacuu	guggcauccucaauuguugaau	66.2	505.2
3684385	uauucuuuuccagaauggccacuc	uguggccagucuggaaaagaaa	51.1	347.8
2746967	aauuugaacguugccaucuauc	uagauggcaacguucaaauucc	49.1	75
3754444	caucguuguuucagaucaucgc	gaugauccgaaacagcgaugcg	33.7	74
1044673	ucuucaacgcuucgccaaucgccu	gcguuugccgggguugaagaug	25.8	1525.2
3780585	cacgccacaccaucuucggcuu	ccgaagaaggugguggcgcgugg	21.3	145.1
3779176	ucuucaaugcuucggcaaucgccu	gcguuugccgggguugaagaug	17.3	551

Table 4.8: Table of predicted miRNAs in *Symbiodinium sp.* The caricature of a pre-miRNA above the table indicates the typical distribution of short reads that map to the mature sequence, stem loop region and star sequence respectively. There are 21 distinct miRNAs with the same mature sequence, star sequence and pre-miRNA sequence (not shown in table) across both libraries. Two of these miRNAs are located in two separate genomic contigs.

Genome annotation	Mature miRNA sequence	Star miRNA sequence	miRDeep2 score
6770795	acccguagauccgaacuugugg	acagguucguauuuauggucc	19925.8
6274679	uaucgaauccgucaaaaagaga	ucuuuuugauggcugcgaaaca	14139.3
6716944	uaucgaauccgucaaaaagaga	ucuuuuugauggcugcgaaaca	14139.3
6746032	ucagggauuguggugaguuaguu	ccagcucauaacagugccuguua	9831.3
6541202	ucagggauuguggugaguuaguu	ccagcucauaacagugccuguua	9830.7
6541203	ucagggauuguggugaguuaguu	ccagcucauaacagugccuguua	9830.7
6773118	aaagaaguacaagugguaggg	cugccacuuguauuuucuuuca	6702.8
6791872	gagguccggaugguuga	auccgcuugauugaccucauu	6348.9
6791872	gagguccggaugguuga	auccgcuucaacgaccucauuu	6313.8
6645066	uaucgauuccgucaaaaagaga	ucuuuuugauggcugcgaaaca	3370.7
6229238	uaugauaucguauccuuugagg	ucaggguuaugaaucauagg	1378.8
6474890	uaugauaucguauccuuugagg	ucaggguuaugaaucauagg	1378.8
6490774	uaugauaucguauccuuugagg	ucaggguuaugaaucauagg	1378.8
6490775	uaugauaucguauccuuugagg	ucaggguuaugaaucauagg	1378.8
6769577	aaguuugagauuugauuuacugaag	cgguugaggauauuuuaaugaucaaagu	1097.6
6785910	ucucugaaaucuccuaagcuauca	aagaguuuagggauuucaggaaa	976.8
6381938	ucaguuccaccaucucaccuaua	aggugagcuguaugaacuuuua	937.2
6625282	ucaguuccaccaucucaccuaua	aggugagcuguaugaacuuuua	937.2
6651661	ucaguuccaccaucucaccuaua	aggugagcuguaugaacuuuua	937.2
6625282	ucaguuccaccaucucaccuaua	aggugagugguaugacuugua	916
6659327	uuccgaugucugugauuuuauuucg	aauaaaauccaggccuuggaga	885.2
6388671	uuccgaugucugugauuuuauuucg	aauaaaauccaggccuuggaga	884.9
6388672	uuccgaugucugugauuuuauuucg	aauaaaauccaggccuuggaga	884.9
6618940	ggaguuuguuguacugugcuauu	ugcauaguguaacaaauuccauc	803.4
6618941	ggaguuuguuguacugugcuauu	ugcauaguguaacaaauuccauc	803.4
6634997	ggaguuuguuguacugugcuauu	ugcauaguguaacaaauuccauc	803.4
6656971	ggaguuuguuguacugugcuauu	ugcauaguguaacaaauuccauc	803.4
6779105	ugggauuaaaaacuucuucggugugg	ucaccgagaauuuuuaaucuga	665.7
6445131	caauguuucggcuuguucccg	ggaacaagccgaaacacugaac	640.6
6786029	caauguuucggcuuguucccg	ggaacaagccgaaacacugaac	640.6
6713016	ucaagucuaggcuggguuaguuu	cuauaccagaauaggcuucag	548.4
6784505	uuuaguuuuccgauauuuuuagg	ugaaaauguuggaaaauuauauc	341.6
6326501	ugaacccagaaccucgaagg	cuucgaggagcuaggguuuaua	291.4
6653489	ugaacccagaaccucgaagg	cuucgaggagcuaggguuuaua	291.4
6374900	ugaaauacucugacggagucagu	gcauuccaucagaauauuuucgcg	242.3
6430146	ugaaauacucugacggagucagu	gcauuccaucagaauauuuucgcg	242.3
6764887	ugucauauccauccaaacgagg	ucuuuucgauggguacgaaaca	233.9
6541717	ugugauuggagacuuuuaucgu	ggugaaagucuucaguuacucu	232.7
6600552	ugugauuggagacuuuuaucgu	ggugaaagucuucaguuacucu	232.7
6718057	ugugauuggagacuuuuaucgu	ggugaaagucuucaguuacucu	232.7
6462057	uguuauaccucagacuucaugc	uugaaagcugaggcauaaucacca	136.6
6693180	uguuauaccucagacuucaugc	uugaaagcugaggcauaaucacca	136.6
6715211	ccgauuugaacaauguuccguuc	cggauuauuguucaaauaag	125.9
6790676	aaauugcuccgaaauacaucuau	cgauguacuucagagcaauuuuu	88.1
6612414	uccacacucaggauguacuagu	uacaacauccugggugugu	74.9
6773945	uuugcuaguugcuuuuguccccguu	agggcaaagguucccagcacgug	73

Genome annotation	Mature miRNA sequence	Star miRNA sequence	miRDeep2 score
6558001	uccagcaccaauguuauuguua	augauaacguugaugcuguaua	68.8
6775082	uucgaucagugucguugacuaacu	uagucacgucacugguuugaauu	63.9
6420714	uuaggcagguaucuaggauc	uccuagauuucuaccuauu	63.6
6468341	uuaggcagguaucuaggauc	uccuagauuucuaccuauu	63.6
6400757	uggcauaagggcagccaccccuu	agggucgcucccuuaugccuca	61.8
6504900	uggcauaagggcagccaccccuu	agggucgcucccuuaugccuca	61.8
6602386	uggcauaagggcagccaccccuu	agggucgcucccuuaugccuca	61.8
6245716	uauauuguacgacucucaucgugu	cggugaaagucgcuucaauaaaca	49.4
6509689	uauauuguacgacucucaucgugu	cggugaaagucgcuucaauaaaca	49.4
6680057	uauauuguacgacucucaucgugu	cggugaaagucgcuucaauaaaca	49.4
6777578	uuaggaauaucgaggacucgcau	ucgagcccuugauauuucagc	40.6
6767213	uaaauucaguguugucagaugu	ucuaacaacacugaaauuacaca	33.9
6739184	ccaacugugacugcaaauuaau	uaauguacaguuacuauugguu	32.1
6524490	ccaacugugacugcaaauuaau	uaauguacaguuacuauugguu	31.9
6524878	ccaacugugacugcaaauuaau	uaauguacaguuacuauugguu	31.9
6706988	ccaacugugacugcaaauuaau	uaauguacaguuacuauugguu	31.9
6246583	acugauauucaccaagugauua	cucauuugcugauuaucagacua	31.3
6450747	acugauauucaccaagugauua	cucauuugcugauuaucagacua	31.3
6583446	acugauauucaccaagugauua	cucauuugcugauuaucagacua	31.3
6773513	aaccagagaaccucagcauuugu	aauuguucagguccuauggauu	23.5
6447759	gaaaaguucgucgaucacucg	gcgugauuuacaaacuuuucu	23.4
6447760	gaaaaguucgucgaucacucg	gcgugauuuacaaacuuuucu	23.4
6635206	gaaaaguucgucgaucacucg	gcgugauuuacaaacuuuucu	23.4
6720692	gaaaaguucgucgaucacucg	gcgugauuuacaaacuuuucu	23.4
6483766	acagccuaauggaccaauguga	ccguuggucaauuagguugau	20.7
6483767	acagccuaauggaccaauguga	ccguuggucaauuagguugau	20.7
6541239	acagccuaauggaccaauguga	ccguuggucaauuagguugau	20.6
6593848	acagccuaauggaccaauguga	ccguuggucaauuagguugau	20.6
6625282	ucaguuccaccaucucaccuac	ggugagcuguaugacuugua	19.2
6785057	gaaguggaggugaauaguggcgg	accacuauucaccgucacugagg	18
6619319	uagcauaacauuguaagagauc	gcucuugcauugcugugcuguc	16.2
6619320	uagcauaacauuguaagagauc	gcucuugcauugcugugcuguc	16.2
6619321	uagcauaacauuguaagagauc	gcucuugcauugcugugcuguc	16.2
6624835	uagcauaacauuguaagagauc	gcucuugcauugcugugcuguc	16.2
6786080	cccaccaauaagcauagccgc	ugucuaugcuucuugaggguu	14.8
6783793	ugugcaagaauuugagucgcugg	agcgacucaaauuucuguaucaca	14.1
6248000	uucgaggaaaugucacuuacg	aagugacauuuccucga	11.1
6416582	uucgaggaaaugucacuuacg	aagugacauuuccucga	11.1
6248000	uucgaggaaaugucacuuacg	aagugacauuuccucga	11
6416582	uucgaggaaaugucacuuacg	aagugacauuuccucga	11
6716075	auauuucaaaguacgcguucuuu	augagcgcgcacuuugaauuauu	10.2
6784141	uauccagagacaaauguuuaauu	agaauauuuggcucugagauauu	10.2
6788906	auauuucaaaguacgcguucuuu	augagcgcgcacuuugaauuauu	10.2
6784016	auauuucaaaguacgcguucuuu	augagcgcgcacuuugaauuauu	10.1

Table 4.9: Table of predicted miRNAs in *S. pistillata*. There are 46 distinct miRNAs with the same mature sequence, star sequence and pre-miRNA sequence (not shown in table) across both libraries.

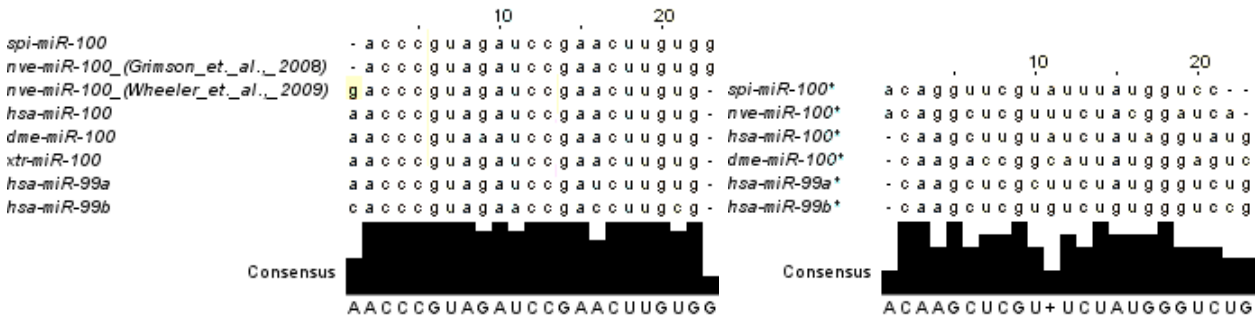


Figure 4.19: Alignment of spi-miR-100 against members of the miR-100 family. The mature sequence is shown on the left, while the star sequence is on the right. The three-letter abbreviations not expounded in the text are "hsa": *H. sapiens*; "dme": *D. melanogaster*; "xtr": *X. tropicalis*. The conflicting nve-miR-100 sequences are from two different publications (Grimson *et al.*, 2008; Wheeler *et al.*, 2009), as denoted in the annotation.

et al. (2009) and Grimson *et al.* (2008) predicted mature miR-100 sequences which are offset by a single nucleotide. From our read data, we had >30,000 reads that exactly match the nve-miR-100 from Grimson *et al.* (2008), but none that matched the alternative version from Wheeler *et al.* (2009) (there were ∼10 reads which were shifted one nucleotide upstream). In the off-chance that the actual spi-miR-100 is one base pair upstream of its current form, the shifted version will be an identical match to hsa-, dme- and xtr-miR-100s as an "A" precedes the current spi-miR-100 sequence.

In humans, miR-100 has been shown to regulate cell differentiation and survival. The underexpression of miR-100 has been noted in cancerous ovarian (Nagaraja *et al.*, 2010) and nasopharyngeal cells (Shi *et al.*, 2010). However, as miRNA-mRNA target recognition depends largely on the miRNA seed sequence (bases 2–7 of the mature miRNA), the targets of hsa-miR-100 and spi-miR-100 will be different due to the one nucleotide offset between the two miRNA sequences. This hypothesis is in agreement with Grimson *et al.* (2008), owing to the match between our miR-100s. Despite the offset, our spi-miR-100 adds to the existing literature documenting the strong conservation of miR-100 amongst metazoans.

On the other hand, spi-miR-2023 (an exact match to nve-miR-2023) might be an Anthozoa-specific miRNA, as it has not been found in any other organism except for *N. vectensis*. Unlike miR-100, no biological function has been discovered for this miRNA.

For both miRNAs, the star sequences are less conserved than the mature miRNA sequences. This is in line with our current understanding of miRNA function, and serves to strongly indicate functional conservation of these two miRNAs in *S. pistillata*.

BLAST searches for both miRNAs in the other coral genome (*A. digitifera*, `http://marinegenomics.oist.jp/genomes/gallery`) did not produce any matches. It is possible that *A. digitifera* lost both miRNAs after diverging from *S. pistillata*, but definite conclusions cannot be drawn at this stage due to the lack of genomic data from other corals. A clearer picture of miRNA evolution within corals will emerge once the "ReFuGe 20/20" initiative gains traction.

4.3.3 Identification of condition-specific short reads in *Symbiodinium sp.*

Using MeV, short reads with similar expression patterns were grouped together. In an effort to reduce chance effects on expression patterns, groups that had less than ten members were automatically discarded from the analysis.

Among the remaining groups, we observed several groups of reads which were strongly overexpressed in a specific condition, but generally underexpressed in the other conditions. One example of such a group is shown in Figure 4.20.

While many groups exhibited condition-specific expressions that were within a standard deviation from the mean, far fewer groups had expressions that were 1.96 standard deviations away from the mean (corresponding to $P < 0.05$). All of these reads are overexpressed ($>$ 1.96 standard deviation) — none of them belong to groups that showed significant condition-specific underexpression. A list of these 457 condition-specific overexpressed reads is in the Appendix (Section 5.3.3).

Interestingly, this list does not contain any of the candidate *Symbiodinium sp.* miRNAs identified in the previous section. It could either mean that none of the candidate miRNAs are involved in stress-response mechanisms, or if it did, it does not involve large changes in the expression of the miRNA.

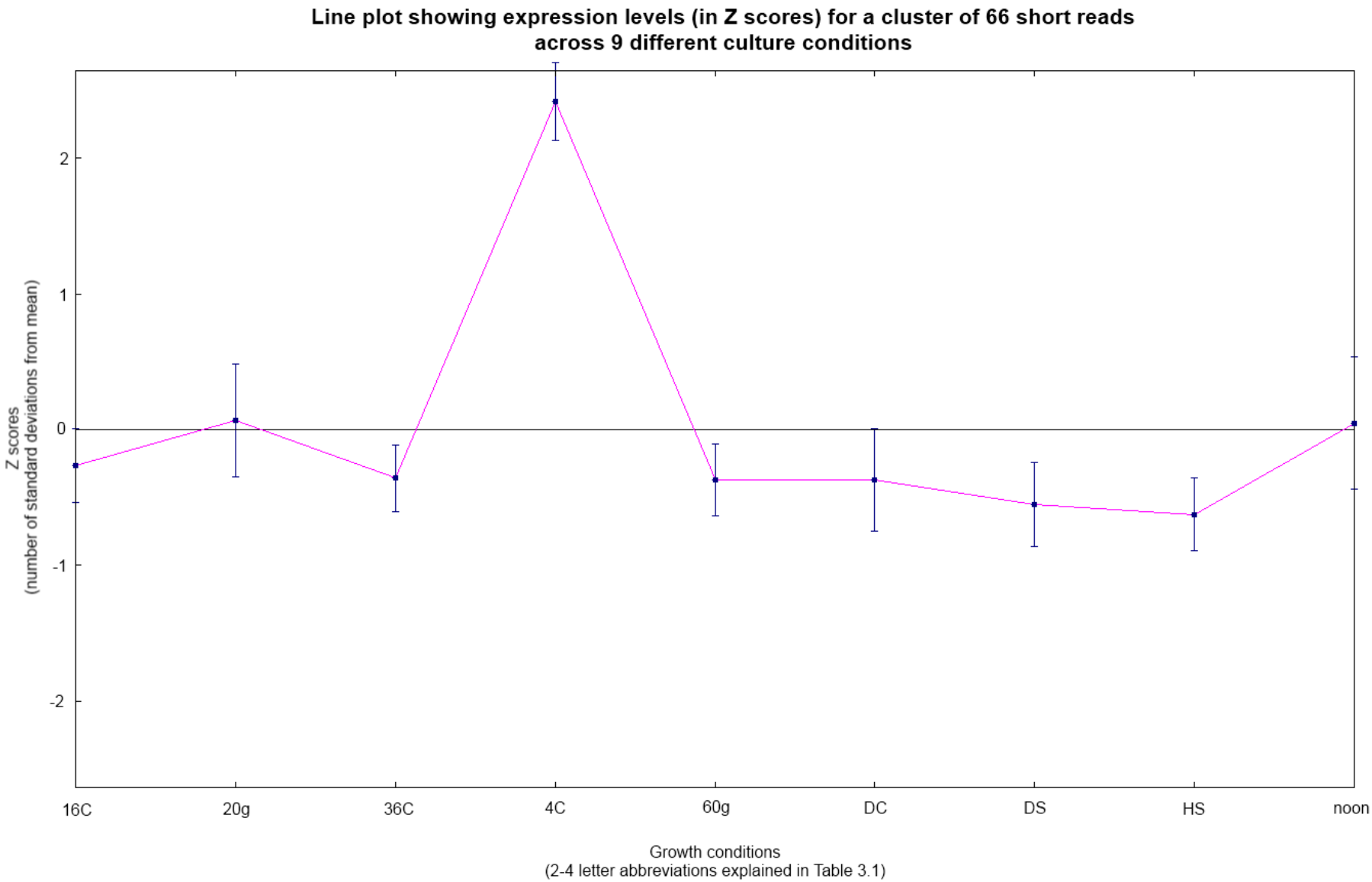


Figure 4.20: A group of 66 short reads showing strong overexpression under extreme cold stress (labelled as "4C"). The expression of each short read under extreme cold stress is ∼2.2 standard deviations above the mean.

Admittedly, due to the lack of biological and technical replicates, reads that exhibit *bona fide* condition-specific expression may well be only a small fraction of the identified reads. In its current form, the list of condition-specific reads is best used as a guide to hone further investigations into identifying reads that are selectively expressed in response to certain stresses, may it be for comparison with future sequence data from other *Symbiodinium* strains, or as a list of candidate reads that could be verified in the wet lab when molecular techniques are adapted to study dinoflagellates.

4.4 Conclusion

With each passing year, advances in our understanding of small RNAs have shown that these small molecules have a big impact in the regulation of gene expression. While the biogenesis and downstream function of the three major sRNA classes — siRNAs, miRNAs and piRNAs — is fairly-well studied in model organisms and shown to be conserved across many diverse eukaryotes, the details behind many other classes of sRNAs discovered in lesser-studied species still remain unclear (take for example the sRNA classes identified in *T. thermophila* and *P. tetraurelia*).

In this chapter, preliminary investigations into the small RNAome of two organisms, the coral *S. pistillata* and its associated dinoflagellate symbiont *Symbiodinium sp.,* were carried out as part of a larger effort that aims to characterise the genome, transcriptome and proteome of those two organisms. An improved understanding of the biology of these organisms will be crucial in devising efforts to slow the imminent destruction of corals (the "rainforest of the sea") worldwide, especially when rates of ocean acidification and global warming show no signs of abating.

I started out by identifying candidate proteins that are homologues of known RNAi proteins from the transcript and protein data of both organisms. The initial protein BLAST searches against known RNAi protein sequences was then supplemented with the alignment of Dicer and Argonaute protein sequences to search for conservation of key residues in the proteins. Based on the high degree of conservation of these key residues in *Symbiodinium sp.*

and *S. pistillata*, the presence of a functional RNAi machinery in both coral and dinoflagellate seems plausible.

As miRNA sequences are endogenously produced, the availability of preliminary versions of the genomes allows for the prediction of miRNAs in both organisms. While none of the miRNAs predicted in *Symbiodinium sp.* matched known sequences, two miRNAs predicted in *S. pistillata* were found to match their respective homologues in *N. vectensis* (a sea anemone). One of the miRNAs, miR-100, is conserved in most model organisms and has been shown to have a role in cell differentiation and survival; the other miRNA, miR-2023, has yet to have function assigned to it.

The availability of condition-specific short read sequences for *Symbiodinium sp.* has made it possible to identify short reads that are over- or under-expressed in specific growth conditions, possibly as a response to the stress faced by the dinoflagellate. A total of 457 reads were found to be exhibit condition-specific overexpression. However, as this list was produced in the absence of replicates, further work is required to verify the identity of the reads and the extent of overrepresentation.

4.5 Further work

As the genome, transcriptome and proteome from *S. pistillata* has yet to be finalised for publication, future refinements to the datasets used in this analysis will lead to changes in the current number of candidate *S. pistillata* homologues of known RNAi proteins (or even new homologues to proteins such as Drosha), and the number of predicted miRNAs. The pipelines used in the current analysis will be re-ran once these datasets are publication-ready.

Based on the presence of functional RNAi machineries in both *Symbiodinium sp.* and *S. pistillata*, we are also interested in identifying functional piRNAs in both organisms. For many vertebrates, flies, and *N. vectensis*, piRNAs have a strong bias for 5'-terminal uracil (5'-U). Analysis of our data reveals a strong enrichment of 5'-U in the 25–30 nt region for *S. pistillata* but not in *Symbiodinium sp.* (see Figure 4.21), indicating that piRNAs is likely to be present in the former but not in the latter. However, in order to further verify the

presence of piRNAs in the libraries, experiments such as comparing periodate-treated short read sequences to non-treated ones need to be carried out. Periodate is a chemical that modifies RNA without a methylation at the terminal 2' oxygen to a form that cannot be sequenced. As piRNAs have this characteristic methylation on the terminal 2' oxygen, they are unaffected by periodate (Grimson *et al.*, 2008).

Lastly, with regards to the post-transcriptional repression of gene expression by miRNAs, we are interested in predicting the mRNA targets of our predicted miRNAs, and to relate changes in miRNA expression to changes in transcript levels of *Symbiodinium sp.* cultured under different stresses. Preliminary scripts have been written to investigate the inverse relationship between miRNA and transcript expressions, but it seems that miRNA expression does not seem to be an important factor behind the changes in transcript expression levels. Due to the lack of experimental data regarding miRNA-mRNA targeting in dinoflagellates, we do not know whether the targeting mechanism is more similar to metazoans or to plants (the latter requiring more exact pairing between the miRNA and mRNA). To compound this, neither do we know whether targeting should be restricted to the 3' UTR or the entire transcript (both have been tested, but both were inconclusive), nor sure that our predicted miRNAs are functional in the first place.

Also, assigning function to these miRNAs is a large project in itself — unfortunately, even though there are many ongoing efforts at studying the biology of *Symbiodinium sp.*, functional genomics has only started to gain traction in the marine science community at present. On a more upbeat note, there are many exciting upcoming projects within the marine sciences (the sequencing of ten coral genomes by ReFuGe 20/20; development and adaptation of new and existing molecular techniques to verify predictions based on sequencing data etc.), and it is hoped that successes from these projects will spark further interest and spur future efforts in understanding these organisms that lie "twenty thousand leagues" under the sea.

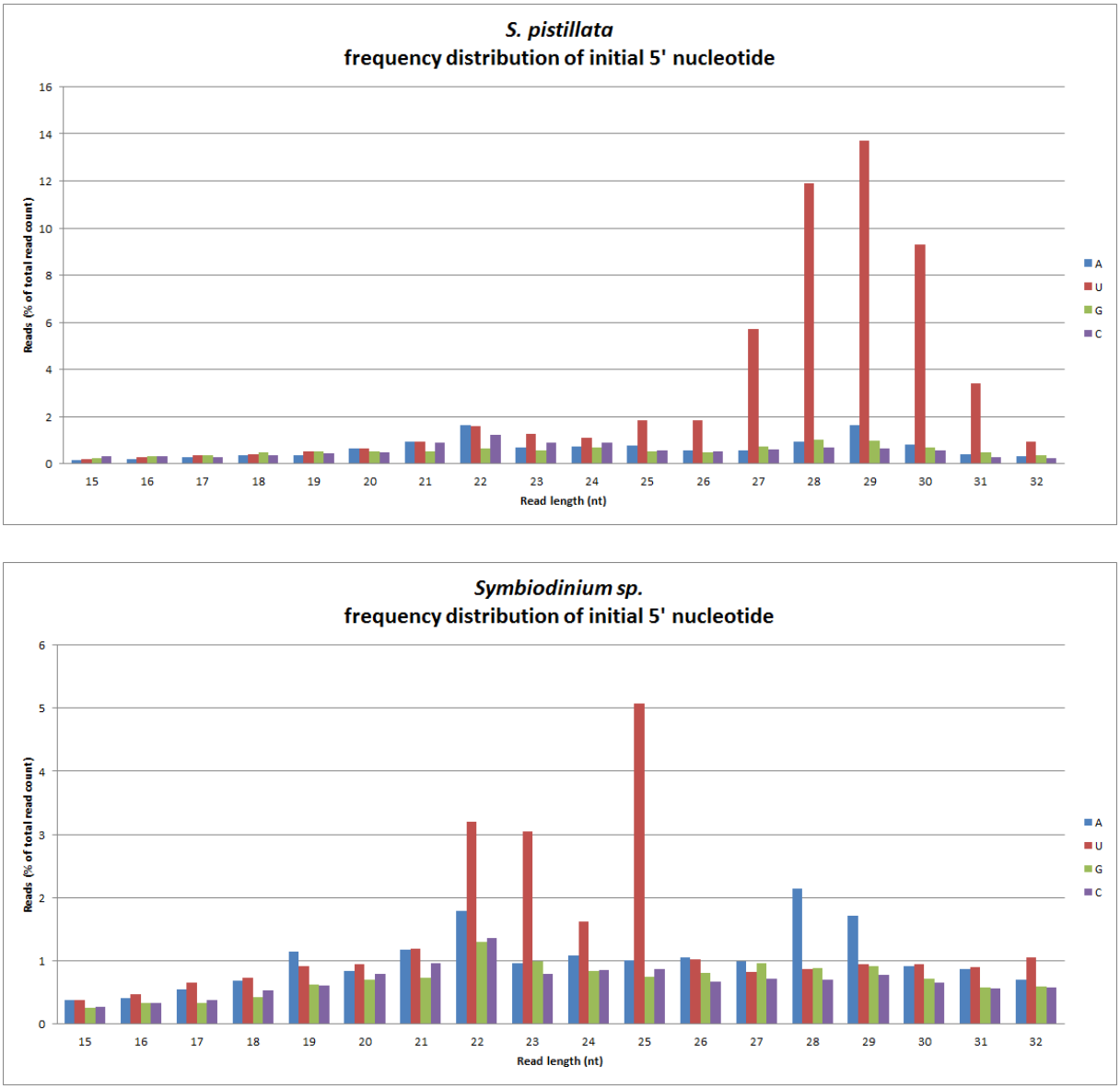


Figure 4.21: Frequency distributions of initial 5' nt in *S. pistillata* (top) and *Symbiodinium sp.* (bottom). There is a significant enrichment of reads containing 5'-U of length 25–30 nt in *S. pistillata*, but not in *Symbiodinium sp.*

Bibliography

Ambros, V., B. Bartel, D. Bartel, C. Burge, J. Carrington, *et al.*, 2003 A uniform system for microRNA annotation. RNA **9**: 277–279.

Aravin, A., M. Lagos-Quintana, A. Yalcin, M. Zavolan, D. Marks, *et al.*, 2003 The small RNA profile during Drosophila melanogaster development. Developmental Cell **5**: 337–350.

Bak, R., 1987 Effects of chronic oil pollution on a Caribbean coral reef. Marine Pollution Bulletin **18**: 534–539.

Baker, A., 2003 Flexibility and specificity in coral-algal symbiosis: diversity, ecology, and biogeography of Symbiodinium. Annual Review of Ecology, Evolution, and Systematics : 661–689.

Bartel, D., 2004 MicroRNAs genomics, biogenesis, mechanism, and function. Cell **116**: 281–297.

Bartel, D., 2009 MicroRNAs: target recognition and regulatory functions. Cell **136**: 215–233.

Bayer, T., M. Aranda, S. Sunagawa, L. Yum, M. DeSalvo, *et al.*, 2012 Symbiodinium Transcriptomes: Genome Insights into the Dinoflagellate Symbionts of Reef-Building Corals. PLoS One **7**: e35269.

Berezikov, E., N. Robine, A. Samsonova, J. Westholm, A. Naqvi, *et al.*, 2011 Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. Genome Research **21**: 203–215.

Bernstein, E., A. Caudy, S. Hammond, and G. Hannon, 2001 Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature **409**: 363–366.

Beyret, E., N. Liu, and H. Lin, 2012 piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. Cell Research **1**: 11.

Biemar, F., R. Zinzen, M. Ronshaugen, V. Sementchenko, J. Manak, *et al.*, 2005 Spatial regulation of microRNA gene expression in the Drosophila embryo. Proceedings of the National Academy of Sciences **102**: 15907–15911.

Bird, A., 1995 Gene number, noise reduction and biological complexity. Trends in Genetics **11**: 94–100.

Blank, R., and R. Trench, 1986 Nomenclature of endosymbiotic dinoflagellates. Taxon : 286–294.

Bohnsack, M., K. Czaplinski, and D. Gorlich, 2004 Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA **10**: 185–191.

Boldt, L., D. Yellowlees, and W. Leggat, 2009 Measuring Symbiodinium sp. gene expression patterns with quantitative real-time PCR. Proceedings of the 11th ICRS : 118–122.

Borchert, G., W. Lanier, B. Davidson, *et al.*, 2006 RNA polymerase III transcribes human microRNAs. Nature Structural and Molecular Biology **13**: 1097–1101.

Brennecke, J., A. Aravin, A. Stark, M. Dus, M. Kellis, *et al.*, 2007 Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell **128**: 1089–1103.

Brennecke, J., D. Hipfner, A. Stark, R. Russell, and S. Cohen, 2003 bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell **113**: 25–36.

Brennecke, J., A. Stark, R. Russell, and S. Cohen, 2005 Principles of microRNA-target recognition. PLoS Biology **3**: e85.

Brodersen, P., L. Sakvarelidze-Achard, M. Bruun-Rasmussen, P. Dunoyer, Y. Yamamoto, *et al.*, 2008 Widespread translational inhibition by plant miRNAs and siRNAs. Science **320**: 1185–1190.

Brown, C., A. Ballabio, J. Rupert, R. Lafreniere, M. Grompe, *et al.*, 1991 A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature **349**: 38–44.

Bruno, J., and E. Selig, 2007 Regional decline of coral cover in the Indo-Pacific: timing, extent, and subregional comparisons. PLoS One **2**: e711.

Bullard, J., E. Purdom, K. Hansen, and S. Dudoit, 2010 Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics **11**: 94.

Bushati, N., and S. Cohen, 2007 microRNA functions. Annu. Rev. Cell Dev. Biol. **23**: 175–205.

Callieri, C., B. Modenutti, C. Queimalinos, R. Bertoni, and E. Balseiro, 2007 Production and biomass of picophytoplankton and larger autotrophs in Andean ultraoligotrophic lakes: differences in light harvesting efficiency in deep layers. Aquatic Ecology **41**: 511–523.

Carmell, M., Z. Xuan, M. Zhang, and G. Hannon, 2002 The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes & Development **16**: 2733–2742.

Carthew, R., and E. Sontheimer, 2009 Origins and Mechanisms of miRNAs and siRNAs. Cell **136**: 642–655.

Caygill, E., and L. Johnston, 2008 Temporal regulation of metamorphic processes in Drosophila by the let-7 and miR-125 heterochronic microRNAs. Current Biology **18**: 943–950.

Celniker, S., L. Dillon, M. Gerstein, K. Gunsalus, S. Henikoff, *et al.*, 2009 Unlocking the secrets of the genome. Nature **459**: 927–930.

Chapman, E., and J. Carrington, 2007 Specialization and evolution of endogenous small RNA pathways. Nature Reviews Genetics **8**: 884–896.

Chapman, J., E. Kirkness, O. Simakov, S. Hampson, T. Mitros, *et al.*, 2010 The dynamic genome of Hydra. Nature **464**: 592–596.

Chen, C., D. Ridzon, A. Broomer, Z. Zhou, D. Lee, *et al.*, 2005 Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Research **33**: e179.

Chen, K., J. Maaskola, M. Siegal, and N. Rajewsky, 2009 Reexamining microRNA site accessibility in Drosophila: a population genomics study. PLoS One **4**: e5681.

Chendrimada, T., R. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, *et al.*, 2005 TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature **436**: 740–744.

Chintapalli, V., J. Wang, and J. Dow, 2007 Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nature Genetics **39**: 715–720.

Chomczynski, P., and N. Sacchi, 1987 Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Analytical Biochemistry **162**: 156.

Chung, W., K. Okamura, R. Martin, and E. Lai, 2008 Endogenous RNA interference provides a somatic defense against Drosophila transposons. Current Biology **18**: 795–802.

Clamp, M., J. Cuff, S. Searle, and G. Barton, 2004 The Jalview Java alignment editor. Bioinformatics (Oxford, England) **20**: 426–427.

Cloonan, N., A. Forrest, G. Kolle, B. Gardiner, G. Faulkner, *et al.*, 2008 Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods **5**: 613–619.

Copois, V., F. Bibeau, C. Bascoul-Mollevi, N. Salvetat, P. Chalbos, *et al.*, 2007 Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality. Journal of Biotechnology **127**: 549–559.

Dasmahapatra, K., J. Walters, A. Briscoe, J. Davey, A. Whibley, *et al.*, 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature **487**: 94–98.

Daxinger, L., and E. Whitelaw, 2012 Understanding transgenerational epigenetic inheritance via the gametes in mammals. Nature Reviews Genetics **13**: 153–162.

Denli, A., B. Tops, R. Plasterk, R. Ketting, and G. Hannon, 2004 Processing of primary microRNAs by the Microprocessor complex. Nature **432**: 231–235.

Diebel, K., A. Smith, and L. van Dyk, 2010 Mature and functional viral miRNAs transcribed from novel RNA polymerase III promoters. RNA **16**: 170–185.

Dlakic, M., 2006 DUF 283 domain of Dicer proteins has a double-stranded RNA-binding fold. Bioinformatics **22**: 2711–2714.

Doench, J., and P. Sharp, 2004 Specificity of microRNA target selection in translational repression. Genes & Development **18**: 504–511.

Domotor, S., and C. D'Elia, 1984 Nutrient uptake kinetics and growth of zooxanthellae maintained in laboratory culture. Marine Biology **80**: 93–101.

Enright, A., B. John, U. Gaul, T. Tuschl, C. Sander, *et al.*, 2003 MicroRNA targets in Drosophila. Genome Biology **5**: R1.

Faehnle, C., and L. Joshua-Tor, 2007 Argonautes confront new small RNAs. Current Opinion in Chemical Biology **11**: 569–577.

Fire, A., S. Xu, M. Montgomery, S. Kostas, S. Driver, *et al.*, 1998 Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature **391**: 806–811.

Flynt, A., N. Liu, R. Martin, and E. Lai, 2009 Dicing of viral replication intermediates during silencing of latent Drosophila viruses. Proceedings of the National Academy of Sciences **106**: 5270.

Friedländer, M., W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, *et al.*, 2008 Discovering microRNAs from deep sequencing data using miRDeep. Nature Biotechnology **26**: 407–415.

Friedländer, M., S. Mackowiak, N. Li, W. Chen, and N. Rajewsky, 2012 miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research **40**: 37–52.

Friedman, R., K. Farh, C. Burge, and D. Bartel, 2009 Most mammalian mRNAs are conserved targets of microRNAs. Genome Research **19**: 92–105.

Frith, M., M. Pheasant, and J. Mattick, 2005 Genomics: The amazing complexity of the human transcriptome. European Journal of Human Genetics **13**: 894–897.

Förstemann, K., M. Horwich, L. Wee, Y. Tomari, and P. Zamore, 2007 Drosophila microRNAs are sorted into functionally distinct argonaute complexes after production by dicer-1. Cell **130**: 287–297.

Galvani, A., and L. Sperling, 2002 RNA interference by feeding in Paramecium. Trends in Genetics **18**: 11–12.

Garnier, O., V. Serrano, S. Duharcourt, and E. Meyer, 2004 RNA-mediated programming of developmental genome rearrangements in Paramecium tetraurelia. Molecular and Cellular Biology **24**: 7370–7379.

Gorton, K., and G. Micklem, 2009 Developmental and spatial regulation of genes in Drosophila melanogaster (unpublished).

Green, E., and A. Bruckner, 2000 The significance of coral disease epizootiology for coral reef conservation. Biological Conservation **96**: 347–361.

Griffiths-Jones, S., 2004 The microRNA registry. Nucleic acids research **32**: D109–D111.

Griffiths-Jones, S., R. Grocock, S. van Dongen, A. Bateman, and A. Enright, 2006 miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Research **34**: D140.

Griffiths-Jones, S., H. Saini, S. van Dongen, and A. Enright, 2008 miRBase: tools for microRNA genomics. Nucleic Acids Research **36**: D154.

Grimson, A., M. Srivastava, B. Fahey, B. Woodcroft, H. Chiang, *et al.*, 2008 Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature **455**: 1193–1197.

Grün, D., Y. Wang, D. Langenberger, K. Gunsalus, and N. Rajewsky, 2005 microRNA target predictions across seven Drosophila species and comparison to mammalian targets. PLoS Computational Biology **1**: e13.

Guillard, R., and J. Ryther, 1962 Studies of marine planktonic diatoms. I. Cyclotella nana Hustedt, and Detonula confervacea (cleve) Gran. Can. J. Microbiol. **8**: 229–239.

Guindon, S., J. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, *et al.*, 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology **59**: 307–321.

Guindon, S., and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology **52**: 696–704.

Haase, A., L. Jaskiewicz, H. Zhang, S. Lainé, R. Sack, *et al.*, 2005 TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. EMBO Reports **6**: 961–967.

Hackenberg, M., M. Sturm, D. Langenberger, J. Falcon-Perez, and A. Aransay, 2009 miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Research **37**: W68–W76.

Han, J., Y. Lee, K. Yeom, J. Nam, I. Heo, *et al.*, 2006 Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell **125**: 887–901.

Hardcastle, T., and K. Kelly, 2010 baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics **11**: 422.

Harrison, P., and C. Wallace, 1990 Reproduction, dispersal and recruitment of scleractinian corals. Ecosystems of the World **25**: 133–207.

Hayashita, Y., H. Osada, Y. Tatematsu, H. Yamada, K. Yanagisawa, *et al.*, 2005 A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. Cancer Research **65**: 9628–9632.

He, L., and G. Hannon, 2004 MicroRNAs: small RNAs with a big role in gene regulation. Nature Reviews Genetics **5**: 522–531.

Heid, C., J. Stevens, K. Livak, and P. Williams, 1996 Real time quantitative PCR. Genome Research **6**: 986–994.

Huang, P., Y. Liu, C. Lee, W. Lin, R. Gan, *et al.*, 2010a DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Research **38**: W385–W391.

Huang, V., Y. Qin, J. Wang, X. Wang, G. Lin, *et al.*, 2010b RNAa is conserved in mammalian cells. PLoS One **5**: e8848.

Huang, Y., L. Ji, Q. Huang, D. Vassylyev, X. Chen, *et al.*, 2009 Structural insights into mechanisms of the small RNA methyltransferase HEN1. Nature **461**: 823–827.

Hughes, T., 1994 Catastrophes, phase shifts, and large-scale degradation of a Caribbean coral reef. Science **265**: 1547–1551.

Hughes, T., A. Baird, D. Bellwood, M. Card, S. Connolly, *et al.*, 2003 Climate change, human impacts, and the resilience of coral reefs. Science **301**: 929–933.

Iida, S., A. Kobiyama, T. Ogata, and A. Murakami, 2008 The D1 and D2 proteins of dinoflagellates: unusually accumulated mutations which influence on PSII photoreaction. Photosynthesis Research **98**: 415–425.

Iwasaki, S., T. Kawamata, and Y. Tomari, 2009 Drosophila Argonaute1 and Argonaute2 employ distinct mechanisms for translational repression. Molecular Cell **34**: 58–67.

Janowski, B., S. Younger, D. Hardy, R. Ram, K. Huffman, *et al.*, 2007 Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. Nature chemical biology **3**: 166–173.

Jaubert, J., 1989 An integrated nitrifying-denitrifying biological system capable of purifying sea water in a closed circuit aquarium. Bull. Inst. Océan. Monaco **5**: 101–106.

Johnson, S., S. Lin, and F. Slack, 2003 The time of appearance of the C. elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. Developmental Biology **259**: 364–379.

Jones, A., and R. Berkelmans, 2010 Potential costs of acclimatization to a warmer climate: growth of a reef coral with heat tolerant vs. sensitive symbiont types. PLoS One **5**: e10437.

Jones-Rhoades, M. W., D. Bartel, and B. Bartel, 2006 MicroRNAs and their regulatory roles in plants. Annual Review of Plant Biology **57**: 19–53.

Karlen, Y., A. McNair, S. Perseguers, C. Mazza, and N. Mermod, 2007 Statistical significance of quantitative PCR. BMC Bioinformatics **8**: 131.

Kertesz, M., N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, 2007 The role of site accessibility in microRNA target recognition. Nature Genetics **39**: 1278–1284.

Kheradpour, P., A. Stark, S. Roy, and M. Kellis, 2007 Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Research **17**: 1919–1931.

Khvorova, A., A. Reynolds, and S. Jayasena, 2003 Functional siRNAs and miRNAs exhibit strand bias. Cell **115**: 209–216.

Kim, V., 2005 MicroRNA biogenesis: coordinated cropping and dicing. Nature Reviews Molecular Cell Biology **6**: 376–385.

Kim, V., 2006 Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. Genes & development **20**: 1993.

Kirino, Y., and Z. Mourelatos, 2007 Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. Nature structural & molecular biology **14**: 347–348.

Knowlton, N., R. Brainard, R. Fisher, M. Moews, L. Plaisance, *et al.*, 2010 Coral reef biodiversity. Life in the World's Oceans: Diversity Distribution and Abundance : 65–74.

Koch, W., 2004 Technology platforms for pharmacogenomic diagnostic assays. Nature Reviews Drug Discovery **3**: 749–761.

Koren, S., Z. Dubinsky, and O. Chomsky, 2008 Induced bleaching of Stylophora pistillata by darkness stress and its subsequent recovery. Proceedings of the 11th International Coral Reef Symposium : 139–143.

Kozomara, A., and S. Griffiths-Jones, 2011 miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research **39**: D152–D157.

Krek, A., D. Grün, M. Poy, R. Wolf, L. Rosenberg, *et al.*, 2005 Combinatorial microRNA target predictions. Nature Genetics **37**: 495–500.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl, 2001 Identification of novel genes coding for small expressed RNAs. Science **294**: 853–858.

Lai, E., 2002 Micro RNAs are complementary to 3'UTR sequence motifs that mediate negative post-transcriptional regulation. Nature Genetics **30**: 363–364.

Lai, E., P. Tomancak, R. Williams, G. Rubin, *et al.*, 2003 Computational identification of Drosophila microRNA genes. Genome Biol **4**: R42.

LaJeunesse, T., 2005 "Species" Radiations of Symbiotic Dinoflagellates in the Atlantic and Indo-Pacific Since the Miocene-Pliocene Transition. Molecular Biology and Evolution **22**: 570–581.

LaJeunesse, T., G. Lambert, R. Andersen, M. Coffroth, and D. Galbraith, 2005 SYMBIODINIUM (PYRRHOPHYTA) GENOME SIZES (DNA CONTENT) ARE SMALLEST AMONG DINOFLAGELLATES1. Journal of Phycology **41**: 880–886.

Lall, S., D. Grün, A. Krek, K. Chen, Y. Wang, *et al.*, 2006 A genome-wide map of conserved microRNA targets in C. elegans. Current Biology **16**: 460–471.

Lau, N., L. Lim, E. Weinstein, and D. Bartel, 2001 An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science **294**: 858–862.

Lee, R., and V. Ambros, 2001 An extensive class of small RNAs in Caenorhabditis elegans. Science **294**: 862–864.

Lee, R., R. Feinbaum, and V. Ambros, 1993 The C. elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to lin-14. Cell **75**: 843–854.

Lee, S., and K. Collins, 2006 Two classes of endogenous small RNAs in Tetrahymena thermophila. Genes & Development **20**: 28–33.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, *et al.*, 2003 The nuclear RNase III Drosha initiates microRNA processing. Nature **425**: 415–419.

Lee, Y., I. Hur, S. Park, Y. Kim, M. Suh, *et al.*, 2006 The role of PACT in the RNA silencing pathway. The EMBO Journal **25**: 522.

Lee, Y., M. Kim, J. Han, K. Yeom, S. Lee, *et al.*, 2004a MicroRNA genes are transcribed by RNA polymerase II. The EMBO Journal **23**: 4051–4060.

Lee, Y., K. Nakahara, J. Pham, K. Kim, Z. He, *et al.*, 2004b Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. Cell **117**: 69–81.

Leggat, W., O. Hoegh-Guldberg, S. Dove, and D. Yellowlees, 2007 Analysis of an EST library from the dinoflagellate (Symbiodinium sp.) symbiont of reef-building corals1. Journal of Phycology **43**: 1010–1021.

Letunic, I., and P. Bork, 2007 Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23**: 127–128.

Letunic, I., and P. Bork, 2011 Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Research **39**: W475–W478.

Leutenegger, C., 2001 The real-time TaqMan PCR and applications in veterinary medicine. Vet Sci Tomorrow **1**: 1–15.

Lewis, B., C. Burge, and D. Bartel, 2005 Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell **120**: 15–20.

Lewis, B., I. Shih, M. Jones-Rhoades, D. Bartel, and C. Burge, 2003 Prediction of mammalian microRNA targets. Cell **115**: 787–798.

Li, J., Z. Yang, B. Yu, J. Liu, and X. Chen, 2005 Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. Current Biology **15**: 1501–1507.

Li, L., S. Okino, H. Zhao, D. Pookot, S. Urakami, *et al.*, 2006 Small dsRNAs induce transcriptional activation in human cells. Science's STKE **103**: 17337.

Li, L., D. Pookot, E. Noonan, R. Dahiya, *et al.*, 2008a MicroRNA-373 induces expression of genes with complementary promoter sequences. Proceedings of the National Academy of Sciences **105**: 1608.

Li, N., A. Flynt, H. Kim, L. Solnica-Krezel, and J. Patton, 2008b Dispatched Homolog 2 is targeted by miR-214 through a combination of three weak microRNA recognition sites. Nucleic Acids Research **36**: 4277–4285.

Li, W., and A. Godzik, 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**: 1658–1659.

Li, X., and R. Carthew, 2005 A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye. Cell **123**: 1267–1277.

Liew, Y., and G. Micklem, 2008 Identification of tissue-specific Drosophila miRNA targets (unpublished).

Lim, L., N. Lau, P. Garrett-Engele, A. Grimson, J. Schelter, *et al.*, 2005 Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature **433**: 769–773.

Lim, L., N. Lau, E. Weinstein, A. Abdelhakim, S. Yekta, *et al.*, 2003 The microRNAs of Caenorhabditis elegans. Genes & Development **17**: 991–1008.

Lin, S., 2006 THE SMALLEST DINOFLAGELLATE GENOME IS YET TO BE FOUND: A COMMENT ON LAJEUNESSE ET AL."SYMBIODINIUM (PYRRHOPHYTA) GENOME SIZES (DNA CONTENT) ARE SMALLEST AMONG DINOFLAGELLATES" 1. Journal of Phycology **42**: 746–748.

Lingel, A., B. Simon, E. Izaurralde, and M. Sattler, 2003 Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. Nature **426**: 465–469.

Livak, K., and T. Schmittgen, 2001 Analysis of relative gene expression data using real-time quantitative PCR and the 2-[Delta][Delta] CT method. Methods **25**: 402–408.

Lund, E., S. Guttinger, A. Calado, J. Dahlberg, and U. Kutay, 2004 Nuclear export of microRNA precursors. Science **303**: 95–98.

MacRae, I., K. Zhou, F. Li, A. Repic, A. Brooks, *et al.*, 2006 Structural basis for double-stranded RNA processing by Dicer. Science **311**: 195.

Markell, D., R. Trench, and R. Iglesias-Prieto, 1992 Macromolecules associated with the cell-walls of symbiotic dinoflagellates. Symbiosis **12**: 19–31.

Marshall, O., 2004 PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics **20**: 2471–2472.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. Journal **17**: pp–10.

Matsui, M., F. Sakurai, S. Elbashir, D. Foster, M. Manoharan, *et al.*, 2010 Activation of LDL receptor expression by small RNAs complementary to a noncoding transcript that overlaps the LDLR promoter. Chemistry & biology **17**: 1344–1355.

Mattick, J., 2007 A new paradigm for developmental biology. Journal of Experimental Biology **210**: 1526–1547.

Moss, E., R. Lee, and V. Ambros, 1997 The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. Cell **88**: 637–646.

Moxon, S., F. Schwach, T. Dalmay, D. MacLean, D. Studholme, *et al.*, 2008 A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics **24**: 2252–2253.

Mulder, N., R. Apweiler, *et al.*, 2007 InterPro and InterProScan: tools for protein sequence classification and comparison. Methods in Molecular Biology (Clifton, NJ) **396**: 59.

Murphy, D., B. Dancis, and J. Brown, 2008 The evolution of core proteins involved in microRNA biogenesis. BMC Evolutionary Biology **8**: 92.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, *et al.*, 2008 The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**: 1344–1349.

Nagaraja, A., C. Creighton, Z. Yu, H. Zhu, P. Gunaratne, *et al.*, 2010 A link between mir-100 and FRAP1/mTOR in clear cell ovarian cancer. Molecular Endocrinology **24**: 447–463.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, *et al.*, 1999 KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research **27**: 29–34.

Okamura, K., J. Hagen, H. Duan, D. Tyler, and E. Lai, 2007 The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell **130**: 89–100.

Okamura, K., A. Ishizuka, H. Siomi, and M. Siomi, 2004 Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. Genes & Development **18**: 1655–1666.

Okamura, K., N. Liu, and E. Lai, 2009 Distinct mechanisms for microRNA strand selection by Drosophila Argonautes. Molecular Cell **36**: 431–444.

Okamura, K., N. Robine, Y. Liu, Q. Liu, and E. Lai, 2011 R2D2 organizes small regulatory RNA pathways in Drosophila. Molecular and Cellular Biology **31**: 884–896.

Oliver, T., and S. Palumbi, 2009 Distributions of stress-resistant coral symbionts match environmental patterns at local but not regional scales. Marine Ecology Progress Series **378**: 93–103.

Pasquinelli, A., B. Reinhart, F. Slack, M. Martindale, M. Kuroda, *et al.*, 2000 Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature **408**: 86–89.

Pastorok, R., and G. Bilyard, 1985 Effects of sewage pollution on coral-reef communities. Marine Ecology Progress Series **21**: 175–189.

Pfaffl, M., 2001 A new mathematical model for relative quantification in real-time RT–PCR. Nucleic Acids Research **29**: e45–e45.

Pochon, X., and R. Gates, 2010 A new Symbiodinium clade (Dinophyceae) from soritid foraminifera in Hawai'i. Molecular Phylogenetics and Evolution **56**: 492.

Pochon, X., T. LaJeunesse, and J. Pawlowski, 2004 Biogeographic partitioning and host specialization among foraminiferan dinoflagellate symbionts (Symbiodinium; Dinophyta). Marine Biology **146**: 17–27.

Pritchard, C., H. Cheng, and M. Tewari, 2012 MicroRNA profiling: approaches and considerations. Nature Reviews Genetics **13**: 358–369.

Putnam, N., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, *et al.*, 2007 Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science **317**: 86–94.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, *et al.*, 2005 InterProScan: protein domains identifier. Nucleic Acids Research **33**: W116–W120.

Rhoades, M., B. Reinhart, L. Lim, C. Burge, B. Bartel, *et al.*, 2002 Prediction of plant microRNA targets. Cell **110**: 513–520.

Rinn, J., M. Kertesz, J. Wang, S. Squazzo, X. Xu, *et al.*, 2007 Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. Cell **129**: 1311–1323.

Romano, S., and S. Cairns, 2000 Molecular phylogenetic hypotheses for the evolution of scleractinian corals. Bulletin of Marine Science **67**: 1043–1068.

Romano, S., and S. Palumbi, 1996 Evolution of scleractinian corals inferred from molecular systematics. Science **271**: 640–642.

Rosic, N., and O. Hoegh-Guldberg, 2010 A method for extracting a high-quality RNA from Symbiodinium sp. Journal of Applied Phycology **22**: 139–146.

Ruby, J., C. Jan, C. Player, M. Axtell, W. Lee, *et al.*, 2006 Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in C. elegans. Cell **127**: 1193–1207.

Ruby, J., A. Stark, W. Johnston, M. Kellis, D. Bartel, *et al.*, 2007 Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Research **17**: 1850–1864.

Ruijter, J., C. Ramakers, W. Hoogaars, Y. Karlen, O. Bakker, *et al.*, 2009 Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Research **37**: e45–e45.

Ruvkun, G., V. Ambros, A. Coulson, R. Waterston, J. Sulston, *et al.*, 1989 Molecular genetics of the Caenorhabditis elegans heterochronic gene lin-14. Genetics **121**: 501–516.

Saeed, A., N. Bhagabati, J. Braisted, W. Liang, V. Sharov, *et al.*, 2006 TM4 Microarray Software Suite. Methods in enzymology **411**: 134–193.

Saeed, A., V. Sharov, J. White, J. Li, W. Liang, *et al.*, 2003 TM4: a free, open-source system for microarray data management and analysis. Biotechniques **34**: 374.

Sambrook, J., and D. Russell, 2001 *Molecular Cloning: A Laboratory Manual, 3rd ed.*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Santiago-Vázquez, L., L. Ranzer, and R. Kerr, 2006 Comparison of two total RNA extraction protocols using the marine gorgonian coral Pseudopterogorgia elisabethae and its symbiont Symbiodinium sp. Electronic Journal of Biotechnology **9**: 598–603.

Schmittgen, T., E. Lee, J. Jiang, A. Sarkar, L. Yang, *et al.*, 2008 Real-time PCR quantification of precursor and mature microRNA. Methods **44**: 31–38.

Schroeder, A., O. Mueller, S. Stocker, R. Salowsky, M. Leiber, *et al.*, 2006 The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Molecular Biology **7**: 3.

Schwarz, D., G. Hutvágner, T. Du, Z. Xu, N. Aronin, *et al.*, 2003 Asymmetry in the assembly of the RNAi enzyme complex. Cell **115**: 199–208.

Sempere, L., E. Dubrovsky, V. Dubrovskaya, E. Berger, and V. Ambros, 2002 The expression of the let-7 small regulatory RNA is controlled by ecdysone during metamorphosis in Drosophila melanogaster. Developmental Biology **244**: 170–179.

Seto, A., R. Kingston, and N. Lau, 2007 The coming of age for Piwi proteins. Molecular Cell **26**: 603–609.

Shendure, J., 2008 The beginning of the end for microarrays? Nature Methods **5**: 585.

Shi, W., N. Alajez, C. Bastianutto, A. Hui, J. Mocanu, *et al.*, 2010 Significance of Plk1 regulation by miR-100 in human nasopharyngeal cancer. International Journal of Cancer **126**: 2036–2048.

Shinzato, C., E. Shoguchi, T. Kawashima, M. Hamada, K. Hisata, *et al.*, 2011 Using the Acropora digitifera genome to understand coral responses to environmental change. Nature **476**: 320–323.

Sievers, F., A. Wilm, D. Dineen, T. Gibson, K. Karplus, *et al.*, 2011 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology **7**.

Smalheiser, N., 2003 EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. Genome Biology **4**: 403.

Sokol, N., and V. Ambros, 2005 Mesodermally expressed Drosophila microRNA-1 is regulated by Twist and is required in muscles during larval growth. Genes & Development **19**: 2343–2354.

Sokol, N., P. Xu, Y. Jan, and V. Ambros, 2008 Drosophila let-7 microRNA is required for remodeling of the neuromusculature during metamorphosis. Genes & Development **22**: 1591–1596.

Song, J., S. Smith, G. Hannon, and L. Joshua-Tor, 2004 Crystal structure of Argonaute and its implications for RISC slicer activity. Science's STKE **305**: 1434.

Stark, A., J. Brennecke, R. Russell, and S. Cohen, 2003 Identification of Drosophila MicroRNA targets. PLoS Biology **1**: E60.

Stat, M., D. Carter, and O. Hoegh-Guldberg, 2006 The evolutionary history of Symbiodinium and scleractinian hosts–Symbiosis, diversity, and the effect of climate change. Perspectives in Plant Ecology, Evolution and Systematics **8**: 23–43.

Stochaj, W., and A. Grossman, 1997 Differences in the protein profiles of cultured and endosymbiotic Symbiodinium sp.(Pyrrophyta) from the anemone Aiptasia pallida (Anthozoa). Journal of Phycology **33**: 44–53.

Stolarski, J., and E. Roniewicz, 2001 Towards a new synthesis of evolutionary relationships and classification of Scleractinia. Journal of Paleontology **75**: 1090–1108.

Taft, R., M. Pheasant, and J. Mattick, 2007 The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays **29**: 288–299.

Taylor, F., M. Hoppenrath, and J. Saldarriaga, 2008 Dinoflagellate diversity and distribution. Biodiversity and Conservation **17**: 407–418.

ten Lohuis, M., and D. Miller, 1998 Genetic transformation of dinoflagellates (Amphidinium and Symbiodinium): expression of GUS in microalgae using heterologous promoter constructs. The Plant Journal **13**: 427–435.

Tomari, Y., and P. Zamore, 2005 Perspective: machines for RNAi. Genes & Development **19**: 517–529.

Trench, R., 1997 Diversity of symbiotic dinoflagellates and the evolution of microalgal-invertebrate symbioses. Proceedings of the 8th International Coral Reef Symposium **2**: 1275–1286.

Turunen, M., T. Lehtola, S. Heinonen, G. Assefa, P. Korpisalo, *et al.*, 2009 Efficient regulation of VEGF expression by promoter-targeted lentiviral shRNAs based on epigenetic mechanism. Circulation Research **105**: 604–609.

Varkonyi-Gasic, E., R. Wu, M. Wood, E. Walton, and R. Hellens, 2007 Protocol: a highly sensitive RT-PCR method for detection and quantification of microRNAs. Plant Methods **3**: 12.

Vasudevan, S., Y. Tong, and J. Steitz, 2007 Switching from repression to activation: microRNAs can up-regulate translation. Science **318**: 1931–1934.

Veldhuis, M., T. Cucci, and M. Sieracki, 1997 Cellular DNA Content of Marine Phytoplankton Using Two New Fluorochromes: Taxonomic and Ecological Implications1. Journal of Phycology **33**: 527–541.

Vella, M., E. Choi, S. Lin, K. Reinert, and F. Slack, 2004 The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. Genes & Development **18**: 132–137.

Veron, J., 2000 Corals of the world. Australian Institute of Marine Science. Townsville, Australia .

Veron, J., D. Odorico, C. Chen, and D. Miller, 1996 Reassessing evolutionary relationships of scleractinian corals. Coral Reefs **15**: 1–9.

Vester, B., and J. Wengels, 2004 LNA (Locked Nucleic Acid): High-Affinity Targeting of Complementary RNA and DNA. Biochemistry **43**: 13233–13241.

Wakefield, T., M. Farmer, and S. Kempf, 2000 Revised description of the fine structure of in situ "zooxanthellae" genus Symbiodinium. The Biological Bulletin **199**: 76–84.

Walker, N., 2002 A technique whose time has come. Science **296**: 557–559.

Wang, T., and M. Brown, 1999 mRNA quantification by real time TaqMan polymerase chain reaction: validation and comparison with RNase protection. Analytical Biochemistry **269**: 198–201.

Wang, Z., M. Gerstein, and M. Snyder, 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics **10**: 57–63.

Waterhouse, A., J. Procter, D. Martin, M. Clamp, and G. Barton, 2009 Jalview Version 2 — a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**: 1189–1191.

Weis, V., S. Davy, O. Hoegh-Guldberg, M. Rodriguez-Lanetty, and J. Pringle, 2008 Cell biology in model systems as the key to understanding corals. Trends in Ecology & Evolution **23**: 369–376.

Weston, A., W. Dunlap, J. Shick, A. Klueter, K. Iglic, *et al.*, 2012 A Profile of an Endosymbiont-enriched Fraction of the Coral Stylophora pistillata Reveals Proteins Relevant to Microbial-Host Interactions. Molecular & Cellular Proteomics **11**: M111–015487.

Wheeler, B., A. Heimberg, V. Moy, E. Sperling, T. Holstein, *et al.*, 2009 The deep evolution of metazoan microRNAs. Evolution & Development **11**: 50–68.

Whitcombe, D., J. Brownie, H. Gillard, D. McKechnie, J. Theaker, *et al.*, 1998 A homogeneous fluorescence assay for PCR amplicons: its application to real-time, single-tube genotyping. Clinical Chemistry **44**: 918–923.

Wightman, B., T. Burglin, J. Gatto, P. Arasu, and G. Ruvkun, 1991 Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. Genes & Development **5**: 1813–1824.

Williamson, V., A. Kim, B. Xie, G. McMichael, Y. Gao, *et al.*, 2012 Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. Briefings in Bioinformatics .

Winter, J., S. Jung, S. Keller, R. Gregory, and S. Diederichs, 2009 Many roads to maturity: microRNA biogenesis pathways and their regulation. Nature Cell Biology **11**: 228–234.

Yang, X., and L. Li, 2011 miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics **27**: 2614–2615.

Yeom, K., Y. Lee, J. Han, M. Suh, and V. Kim, 2006 Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. Nucleic Acids Research **34**: 4622–4629.

Yi, R., Y. Qin, I. Macara, and B. Cullen, 2003 Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes & Development **17**: 3011–3016.

Zdobnov, E., and R. Apweiler, 2001 InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**: 847–848.

Zeng, Y., and B. Cullen, 2004 Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. Nucleic Acids Research **32**: 4776–4785.

Zerbino, D., and E. Birney, 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research **18**: 821–829.

Chapter 5

Appendix

5.1 Additional data from Chapter 2

5.1.1 Significant tissue-miRNA couples from Liew and Micklem (2008)

miRNA	Tissue	Enriched?	Corrected p value
miR-277	brain	+	0.0100
miR-92b	adult_carcass	-	0.0125
miR-92b	head	-	0.0128
miR-277	testis	+	0.0139
let-7	ovary	-	0.0165
miR-11	adult_carcass	-	0.0216
miR-12	male_acc_glands	+	0.0298
miR-1013	ovary	-	0.0303
miR-8	fat_body_larval	+	0.0366
miR-277	thoracicoabdominal_ganglion	+	0.0435
miR-11	hindgut	-	0.0460

Table 5.1: List of significant tissue-miRNA couples for downregulated transcripts. There are 11 of these couples with P values of < 0.05 (corrected). For more than half of these couples, the proportion of downregulated transcripts targeted by the miRNA is smaller than expected (the "-" signs), which runs contrary to the expected mechanism of miRNA action.

miRNA	tissue	enriched?	corrected p value
miR-11	head	+	0.0010
miR-79	ovary	+	0.0017
let-7	ovary	+	0.0018
miR-11	adult_carcass	+	0.0020
miR-310	head	+	0.0021
miR-124	hindgut	+	0.0022
miR-1013	ovary	+	0.0050
miR-11	hindgut	+	0.0051
miR-308	adult_carcass	+	0.0058
miR-280	ovary	+	0.0059
miR-4	ovary	+	0.0062
miR-184*	ovary	-	0.0062
miR-184*	midgut	+	0.0066
miR-124	tubule_larval	+	0.0070
miR-277	ovary	+	0.0071
miR-210	ovary	+	0.0074
miR-1013	fat_body_larval	+	0.0108
miR-2b	head	+	0.0110
miR-2a	head	+	0.0112
miR-315	testis	+	0.0114
miR-287	adult_carcass	+	0.0132
miR-1	ovary	+	0.0134
miR-92b	adult_carcass	+	0.0135
miR-iab-4-3p	ovary	+	0.0150
miR-308	tubule_larval	+	0.0152
miR-92b	head	+	0.0153
miR-11	crop	+	0.0160
miR-311	head	+	0.0161
miR-1013	tubule_larval	+	0.0164
miR-2c	head	+	0.0166
miR-308	head	+	0.0171
miR-310	tubule_larval	+	0.0181
miR-277	male_acc_glands	+	0.0208
miR-312	head	+	0.0223
miR-284	brain	-	0.0255
miR-284	thoracicoabdominal_ganglion	-	0.0257
miR-308	hindgut	+	0.0264
miR-iab-4-5p	fat_body_larval	+	0.0267
miR-317	tubule	+	0.0272
miR-92b	tubule_larval	+	0.0308
miR-277	fat_body_larval	+	0.0311
miR-1017	brain	-	0.0311
miR-13a	head	+	0.0314
miR-311	thoracicoabdominal_ganglion	+	0.0315
miR-210	male_acc_glands	+	0.0340
miR-310	midgut	+	0.0376
miR-13b	head	+	0.0409
miR-311	tubule_larval	+	0.0410
miR-92a	head	+	0.0412
miR-1010	ovary	-	0.0444
miR-309	male_acc_glands	-	0.0465
miR-2b	hindgut	+	0.0470
miR-4	midgut	+	0.0473
miR-263b	male_acc_glands	-	0.0473

Table 5.2: List of significant tissue-miRNA couples for upregulated transcripts. There are 54 of these couples with P values of < 0.05 (corrected). For most of these couples, the proportion of upregulated transcripts targeted by the miRNA is larger than expected (the "+" signs), which runs contrary to the expected mechanism of miRNA action.

5.2 Additional data from Chapter 3

5.2.1 Full protocol for RNA extractions from *Symbiodinium sp.* cells

Sample collection (used in all extraction methods)

1. Estimate cell density of culture with a haemocytometer.
2. Take 50 ml of sample from the *Symbiodinium sp.* culture, bearing in mind that the sample will be divided into individual Eppendorf tube that each contain 10^6–10^7 cells (so that it does not overload the filters from the commercial kits).
3. Centrifuge at 10,000g for 10 mins.
4. Decant supernatant. Wash pellet at the bottom of the 50 ml Falcon tube with ∼20 ml of MilliQ water. Centrifuge at 10,000g for 10 mins.
5. Decant supernatant and keep Falcon tube inverted for 1–2 minutes to dry pellet. Snap-freeze pellet at -80 °C in liquid nitrogen.

Phenol-chloroform RNA extraction from samples homogenised with bead-beater

1. Add 1.1 ml of QIAzol (Qiagen) into Falcon tube. If the pellet has more than 10^6–10^7 cells, add in additional portions of 1.1 ml QIAzol for every extra planned subdivisions of the pellet (ie. to divide the pellet across five tubes, add 5.5 ml of QIAzol)
2. Vortex the Falcon tube until the pellet disappears and sample is well-mixed.
3. Pipette 1 ml portions of this mixture into 2 ml screw-cap tubes.
4. Add in ∼0.3g glass beads.
5. Homogenise samples in bead beater of choice.

6. Briefly chill the vigorously-shaken tubes in ice to bring it back down to room temperature, and then add 0.2 ml chloroform (0.2 volumes of initial QIAzol, as per manufacturer's protocol). Shake tubes vigorously for 15 seconds and incubate at room temperature for 2–3 minutes.

7. Centrifuge at 12,000g for 15 minutes at 4 °C. Remove upper phase to a new RNase-free tube, being careful not to touch the interface. Discard tube with beads in the lower phase and interface.

8. Add 0.5 ml isopropanol (0.5 volumes of initial QIAzol) to precipitate the RNA. Incubate at room temperature for 10 minutes and then centrifuge at 12,000g for 15 minutes at 4 °C.

9. Wash pellet with 1 ml 75% ethanol (diluted with RNase-free water), then spin for 7,500g for 5 minutes at 4 °C. If the pellet remains floating, centrifuge it at 12,000g for an additional 5 minutes at 4 °C.

10. Remove the supernatant and air-dry the pellet.

11. Resuspend pellet in 100 µl RNase-free water (this is because the RNeasy Mini Kit clean-up step requires the initial volume of the "dirty" sample to be 100 µl).

Note that step 4 and 5 is left intentionally vague, as we varied the bead size, shaking speed and shaking time used in the homogenisation step. Also, to measure the concentration and purity of "pre-wash" samples, 10 µl of sample was taken out into fresh tubes after step 11.

RNA clean up with RNeasy Mini Kit (Qiagen)

12. Adjust the sample to a volume of 100 µl with RNase-free water. Add 350 µl Buffer RLT, and mix well.

13. Add 250 µl ethanol (96–100%) to the diluted RNA, and mix well by pipetting. Do not centrifuge. Proceed immediately to step 3.

14. Transfer the sample (700 µl) to an RNeasy Mini spin column placed in a 2 ml collection tube. Close the lid gently, and centrifuge for 15 s at > 8,000g (> 10,000 rpm). Discard the flow-through. Note: After centrifugation, carefully remove the RNeasy spin column from the collection tube so that the column does not contact the flow-through. Be sure to empty the collection tube completely.

15. Add 500 µl Buffer RPE to the RNeasy spin column. Close the lid gently, and centrifuge for 15 s at > 8,000g (> 10,000 rpm) to wash the spin column membrane. Discard the flow-through. Reuse the collection tube in step 5.

16. Add 500 µl Buffer RPE to the RNeasy spin column. Close the lid gently, and centrifuge for 2 min at > 8,000g (> 10,000 rpm) to wash the spin column membrane. The long centrifugation dries the spin column membrane, ensuring that no ethanol is carried over during RNA elution. Residual ethanol may interfere with downstream reactions.

NOTE: After centrifugation, carefully remove the RNeasy spin column from the collection tube so that the column does not contact the flow-through. Otherwise, carryover of ethanol will occur.

17. Optional: Place the RNeasy spin column in a new 2 ml collection tube (supplied by kit), and discard the old collection tube with the flow-through. Close the lid gently, and centrifuge at full speed for 1 min.

NOTE: Perform this step to eliminate any possible carryover of Buffer RPE, or if residual flow-through remains on the outside of the RNeasy spin column after step 5.

18. Place the RNeasy spin column in a new 1.5 ml collection tube (supplied by kit). Add 30–50 µl RNase-free water directly to the spin column membrane. Close the lid gently, and centrifuge for 1 min at > 8,000g (> 10,000 rpm) to elute the RNA.

mirVana (Ambion) kit-based RNA extraction from samples homogenised in mortar and pestle

1. Pre-chill mortar, pestle and a metal spatula with liquid nitrogen.

2. Dislodge pellet from base of Falcon tube by tapping the tube against the benchtop. Use the metal spatula to scrape pellet fragments that remains clinging to the Falcon tube into the pestle.

3. (Optional) Add 0.3–0.5 g of glass beads (0.1 mm) into the pestle to aid the grinding process.

4. Grind samples thoroughly by pressing hard on the mortar. Keep samples chilled by pouring liquid nitrogen into the pestle occasionally.

5. Scrape the homogenised sample into a 1.5 ml or 2 ml Eppendorf tube.

6. Add 1 ml of Lysis/Binding Buffer and 100 µl of miRNA Homogenate Additive (both from mirVana kit), and mix well by vortexing or inverting the tube several times.

7. Leave the mixture on ice for 10 minutes.

8. If glass beads were added to the pestle (optional step 3), centrifuge the mixture at 13,000g for 5 minutes to pellet the glass beads.

   NOTE: In our experiments, we created two technical duplicates from the same homogenised sample by pipetting 500 µl into two separate tubes.

The following part is taken from Ambion's mirVana miRNA Isolation Kit Protocol, and edited for clarity.

9. Add a volume of Acid-Phenol:Chloroform that is equal to the lysate volume before addition of the miRNA Homogenate Additive. For example, if the original lysate volume was 300 µl, add 300 µl Acid-Phenol:Chloroform.

IMPORTANT: Be sure to withdraw from the bottom phase in the bottle of Acid-Phenol:Chloroform, because the upper phase consists of an aqueous buffer.

10. Vortex for 30–60 sec to mix.

11. Centrifuge for 5 min at maximum speed (10,000g) at room temperature to separate the aqueous and organic phases. After centrifugation, the interphase should be compact; if it is not, repeat the centrifugation.

12. Carefully remove the aqueous (upper) phase without disturbing the lower phase, and transfer it to a fresh tube. Note the volume removed.

13. Add 1/3 volume of 100% ethanol to the aqueous phase recovered from the organic extraction (e.g. add 100 µl 100% ethanol to 300 µl aqueous phase). Mix thoroughly by vortexing or inverting the tube several times.

14. For each sample, place a Filter Cartridge into one of the Collection Tubes supplied. Pipette the lysate/ethanol mixture (from the previous step) onto the Filter Cartridge. Up to 700 µl can be applied to a Filter Cartridge at a time. For sample volumes greater than 700 µl, apply the mixture in successive applications to the same filter.

15. Centrifuge for 15 sec to pass the mixture through the filter. Centrifuge at 10,000g (typically 10,000 rpm). Spinning harder than this may damage the filters. Alternatively, vacuum pressure can be used to pull samples through the filter.

16. Collect the filtrate. If the lysate/ethanol mixture is >700 µl, transfer the flow-through to a fresh tube, and repeat until all of the lysate/ethanol mixture is through the filter. Pool the collected filtrates if multiple passes were done, and measure the total volume of the filtrate.

At this point, the filter contains a RNA fraction that is depleted of small RNAs, while the filtrate is enriched for small RNAs. Steps 17–26 describes the extraction of the RNA fraction that is enriched for small RNAs.

17. Add 2/3 volume room temperature 100% ethanol to filtrate (i.e. flow-through). For example, if 400 µl of filtrate is recovered, add 266 µl 100% ethanol. Mix thoroughly.

18. For each sample, place a Filter Cartridge into one of the Collection Tubes supplied. Pipette the filtrate/ethanol mixture (from the previous step) onto a second Filter Car-

tridge. Up to 700 µl can be applied to a Filter Cartridge at a time. For sample volumes greater than 700 µl, apply the mixture in successive applications to the same filter.

19. Centrifuge for 15 sec to pass the mixture through the filter. Centrifuge at 10,000g (typically 10,000 rpm). Spinning harder than this may damage the filters. Alternatively, vacuum may be used to pass samples through the filter.

20. Discard the flow-through, and repeat until all of the filtrate/ethanol mixture is through the filter. Reuse the Collection Tube for the washing steps.

21. Apply 700 µl miRNA Wash Solution 1 (working solution mixed with ethanol) to the Filter Cartridge and centrifuge for 5–10 sec or use vacuum to pass the solution through the filter. Discard the flow-through from the Collection Tube, and replace the Filter Cartridge into the same Collection Tube.

22. Apply 500 µl Wash Solution 2/3 (working solution mixed with ethanol) and draw it through the Filter Cartridge as in the previous step.

23. Repeat with a second 500 µl aliquot of Wash Solution 2/3.

24. After discarding the flow-through from the last wash, replace the Filter Cartridge in the same Collection Tube and spin the assembly for 1 min to remove residual fluid from the filter.

25. Transfer the Filter Cartridge into a fresh Collection Tube (provided with the kit). Apply 100 µl of pre-heated (95 °C) Elution Solution (0.1 mM EDTA in nuclease-free water) to the center of the filter, and close the cap. Spin for 20–30 sec at maximum speed to recover the RNA.

26. Collect the eluate (which contains the RNA) and store it at -20 °C or colder.

Steps 27–32 detail the recovery of total RNA that has been selectively depleted of small RNA, which is on the filter after step 16. Steps 27–32 is an exact copy of steps 21–26 (ie. recovering cleaned-up RNA trapped by the filter cartridge).

27. Apply 700 µl miRNA Wash Solution 1 (working solution mixed with ethanol) to the Filter Cartridge and centrifuge for 5–10 sec or use a vacuum to pull the solution through the filter. Discard the flow-through from the Collection Tube, and replace the Filter Cartridge into the same Collection Tube.

28. Apply 500 µl Wash Solution 2/3 (working solution mixed with ethanol) and draw it through the Filter Cartridge as in the previous step.

29. Repeat with a second 500 µl aliquot of Wash Solution 2/3.

30. After discarding the flow-through from the last wash, replace the Filter Cartridge in the same Collection Tube and spin the assembly for 1 min to remove residual fluid from the filter.

31. Transfer the Filter Cartridge into a fresh Collection Tube (provided with the kit). Apply 100 µl of pre-heated (95 °C) Elution Solution to the center of the filter, and close the cap. Spin for 20–30 sec at maximum speed to recover the RNA.

32. Collect the eluate (which contains the RNA) and store it at -20 °C or below.

Note that the eluate collected at step 26 cannot be subjected to a downstream wash with RNeasy Mini Kit (Qiagen), as the filters used in the RNeasy Kit traps RNA of length < 200 bp. Ethanol precipitation should be used instead.

5.3 Additional data from Chapter 4

5.3.1 List of non-default program parameters

5.3.1.1 Cutadapt-1.0 (Martin, 2011)

Program parameters for Cutadapt:

Option	Value	Function
-a	(see below)	Sequence of the 3' adapter
-g	(see below)	Sequence of the 5' adapter
-O	4	Minimum overlap length. Reads are not modified if overlap between read and adapter < 4 bases.

For all FASTQ files processed, the minimum overlap length (-O) was increased to 4 (from a default setting of 3) to reduce the possibility of sequences having the first three nucleotides of the adapter trimmed off by chance (1/64 to 1/256).

For FASTQ files from libraries created using the Small RNA Sample Prep Kit: -a ATCTCGTATGCCGTCTTCTGCTTG -g CGACAGGTTCAGAGTTCTACAGTCCGACGATC

For FASTQ files from libraries created using TruSeq Small RNA Sample Prep Kit: -a TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG -g AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGATC

5.3.1.2 PhyML (Guindon and Gascuel, 2003; Guindon *et al.*, 2010)

The only non-default program parameter for PhyML is:

Option	Value	Function
-b	-1	Bootstraps using aLRT (approximate likelihood ratio test)

5.3.1.3 cd-hit-est (Li and Godzik, 2006)

Non-default program parameters for cd-hit-est are:

Option	Value	Function
-T	6	Multithreaded mode, 6 threads used
-M	0	No upper limit to the amount of memory used by program
-g	1	Guaranteed best match (at expense of speed)
-r	0	Only align in the +/+ direction (RNA library is directional)
-n	4	Word size suited for clustering short reads
-c	0.8	Cluster reads if the shorter sequence has >80% identity to the longer, "representative sequence" of the cluster
-d	30	Lengthens maximum annotation length (for downstream scripts)

5.3.2 Protein sequences for RNAi machinery

5.3.2.1 Dicers

>maker-6213567_STYPI

WHAYNLTVQHEPCDCEVADGGVGVDGTVTRNRAYKRKFPTLLENNFPVCGEPCYVYSLTM

ELTKPWNFERALRSRSKAKATSDSYSIGVLSRKPLPKLPCMAIFDRAGEVTVTVTECCPQ

SIFLTADQLSLLQRFTEYVFKEIARPKKESSATFLSFDPTIAFSGYYILFLKDASSSRSL

AGLSSNNHKEVAFDFMFSLEDKLGCFKNPVYSGPPDITDAEIFRDTVVTATYNEKRSHYY

VADICYDLSPSDPFPNIEVAGTFAEYVKIRYDVEVSLDQPMLDVDHTSSRLNFLLPKYEN

FKGQHPRIPEKNSKRSRKSKVYLIPELCSIHPVPGHLWRQLSNLPAVLYRIESLLVAEEL

RCWVVRDLGIGVVDWPKDVPLPPVTVGETMGEEILPAVGSSAVESNGKRSSQVALVNLTS

PFSAQLPEVLMDLKISNQVELTEKSAAKSLAGGLSNGARESCLSMSSETFLVEHSCAETE

CKPLFRPAPHQGLLIPSKEKEDFWPNKLSDLNVPLPTERRAFDCDSIPANDKESEVSAVS

IWTDPFLSNLYRTCGPPSSLILRALTTTLAGDVFSLERLEVLGDSFVKYENEGVLSFLRG

IKVSNRQLFYLARQRGLPSYMFTRMFNPLVNWLPPGFYLDDDSNAENSGYEHRRLISDTF

LDEVDDEEDDESIFAETESSGYNSDCRQGESPDVQLNSYLHVCCSDKSIADCTEALIGAF

LLCFGLDGAFKFLEWLGMEIARDEKDENLVQKPSSDDSARDAVPKPASTMCHHYTSHVPQ

QASETLQENLSLLDDETVPSSSIDGLSSEHESTVDEFKDVEEALHYRFHDKSLLMQAFTH

SSLPGDYNSVRNSYEQLEFLGDALLDFLVTRYLYVNHRHMSPGELTDLRTALVNNYSFAV

LAVKLGFPRHLRSCSPQLFGMVNKFMVKLKEKERKHATTHRKNDEIYSSVFVMETEGRDD
PEQVEIPKVLGDLFEAVAGAMYLDCGADIE
>maker-6462987_STYPI
MAAAYEACIVLHRNGELDDNLRPKRSLSDDESELEEEEEAAAAEPGGNESKPGSKKRKRQ
YTRKVPEVFVGSLVQDGQSQFLTTITIQVTKLIQPQEFVISERLFLNHTCTFGMLTNKRV
PRMRSFKLYPNYGEVTVSLQCHTSPLRRNNTKQDLDILREFHLFLFSHVLRSPVEFTPES
MDGYFIVMLKDSGRSVDFDRMREELSRAHKQSRGNEPIVEDAVVTKNYVQCSQKYAVLKV
RGDLTPLSPFPEKSQGNTYEEYFRRKYGEMGRIADKNQPLVEAKELSGRVNFLVDRKVGT
KHKRQVIYLVPELCNQIPIRASLLSVSQILPSVLQRVTSLLLVADLKTMVAGEGDTTDES
SLPPIGAEEGSHRPNQAGEVEDDGLDLEDSDDAESEFPGLFSDIRSMFRYIPSSVNHPNS
SLVLQAVTTTHSGDAFNLERLEMLGDSFLKLAVSLHLFCTYSDKDEGKLTRRKTNQISNL
ALYRAATKKSLGEYLQSTQLARDVWCPTGCQFGDVPPNRTTEEREKRKKRKLTATEHASQ
KCITQMIGDKSIADSIEALIGAYLISDGYLGALRFMKFLGLKILPEVDPNDGIESLAKNS
KSGCYARFWPDQTKVTATQGKGDVVFRLTSGLENFESNSILYKFQQKLYLVEALTHASYH
ENRATPSYQRLEFLGDALLDFLVTQHLYFRHVNLSPGDLTDIRQAL
>maker-6574134_STYPI
TLQTSKEGKKICKSSATQLHQVQDVYCNFTGINGSRDKNCSHIMPEYLTKLLNIFGKFIG
LESDPEKRLCAIVFVEKRYTALILSEQINQAAKLNHDLSFVKSNFVTGHGTGGKVNFSSE
TEMNFKKQEEVLRKFRRHEFNVLIATSVVEEGLDVPKCNVVCCFDFPKNFRSYVQSKGRA
RARDSNYYMLVPQELEGEKENDLEILREIEKILFKRCHDRTQPSMRECIESLDTDALQPY
VPVDGGGACVTMSNSISILSKYCSKLPGDRFTQPTPVYKIEEIGKDSYRCKLTLPMNCQL
REDIIGEPMRSKKYAKMAAAYKACIVLYRNGELDDNLRPKRSLSDDESELEEEEEAAAAE
PGGNESKPGSKKRKRQYTRKVPEVFVGSLVQDGQSQFLTTITIQVTKLIQPQEFVISERL
FLNHTCTFGMLTNKRVPRMRSFKLYPNYGEVTVSLQCHTSPLRRNNTKQDLDILREFHLF
LFSHVLRSPVEFTPESMDGYFIVMLKDSGRSVDFDRMREELSRAHKQSRGNEPIVEDAVV
TKNYVQCSQKYAVLKVRGDLTPLSPFPEKSQGNTYEEYFRRKYGEMGRIADKNQPLVEAK
ELSGRVNFLVDRKVGTKHKRQVIYLVPELCNQIPIRASLLSVSQILPSVLQRVTSLLLVA
DLKTMVAGEGDTTDESSLPPIGAEEGSHRPNQAGEVEDDGLDLEDSDDAESEFPGLFSDI

RSMFRYIPSSVNHPNSSLVLQAVTTTHSGDAFNLERLEMLGDSFLKLAVSLHLFCTYSDK
DEGKLTRRKTNQISNLALYRAATKKSLGEYLQSTQLARDVWCPTGCQFGDVPPNRTTEER
EKRKKRKLTATEHASQKCITQMIGDKSIADSIEALIGAYLISDGYLGALRFMKFLGLKIL
PEVDPNDGIESLAKNSKSGCYARFWPDQTKVTATQGKGDVVFRLTSGLENFESNSILYKF
QQKLYLVEALTHASYHENRATPSYQRLEFLGDALLDFLVTQHLYFRHVNLSPGDLTDIRQ
A

>maker-6582237_STYPI

WHAYNLTVQHEPCDCEVADGGVGVDGTVTRNRAYKRKFPTLLENNFPVCGEPCYVYSLTM
ELTKPWNFERALRSRSKAKATSDSYSIGVLSRKPLPKLPCMAIFDRAGEVTVTVTECCPQ
SIFLTADQLSLLQRFTEYVFKEIARPKKESSATFLSFDPTIAFSGYYILFLKDASSSRSL
AGLSSNNHKEVAFDFMFSLEDKLGCFKNPVYSGPPDITDAEIFRDTVVTATYNEKRSHYY
VADICYDLSPSDPFPNIEVAGTFAEYVKIRYDVEVSLDQPMLDVDHTSSRLNFLLPKYEN
FKGQHPRIPEKNSKRSRKSKVYLIPELCSIHPVPGHLWRQLSNLPAVLYRIESLLVAEEL
RCWVVRDLGIGVVDWPKDVPLPPVTVGETMGEEILPAVGSSAVESNGKRSSQVALVNLTS
PFSAQLPEVLMDLKISNQVELTEKSAAKSLAGGLSNGARESCLSMSSETFLVEHSCAETE
CKPLFRPAPHQGLLIPSKEKEDFWPNKLSDLNVPLPTERRAFDCDSIPANDKESEVSAVS
IWTDPFLSNLYRTCGPPSSLILRALTTTLAGDVFSLERLEVLGDSFVKYENEGVLSFLRG
IKVSNRQLFYLARQRGLPSYMFTRMFNPLVNWLPPGFYLDDDSNAENSGYEHRRLISDTF
LDEVDDEEDDESIFAETESSGYNSDCRQGESPDVQLNSYLHVCCSDKSIADCTEALIGAF
LLCFGLDGAFKFLEWLGMEIARDEKDENLVQKPSSDDSARDAVPKPASTMCHHYTSHVPQ
QASETLQENLSLLDDETVPSSSIDGLSSEHESTVDEFKDVEEALHYRFHDKSLLMQAFTH
SSLPGDYNSVRNSYEQLEFLGDALLDFLVTRYLYVNHRHMSPGELTDLRTALVNNYSFAV
LAVKLGFPRHLRSCSPQLFGMVNKFMVKLKEKERKHATTHRKNDEIYSSVFVMETEGRDD
PEQVEIPKVLGDLFEAVAGAMYLDCGADIEMITTQSILPCQPFGDCWR

>maker-6743944_STYPI

MGRIADKNQPLVEAKELSGRVNFLVDRKVGTKHKRQVIYLVPELCNQIPIRASLLSVSQI
LPSVLQRVTSLLLVADLKTMVAGEGDTTDESSLPPIGAEEGSHRPNQAGEVEDDGLDLED
SDDAESEFPGLFSDIRSMFRYIPSSVNHPNSSLVLQAVTTTHSGDAFNLERLEMLGDSFL

KLAVSLHLFCTYSDKDEGKLTRRKTNQISNLALYRAATKKSLGEYLQSTQLARDVWCPTG
CQFGDVPPNRTTEEREKRKKRKLTATEHASQKCITQMIGDKSIADSIEALIGAYLISDGY
LGALRFMKFLGLKILPEVDPNDGIESLAKNSKSGCYARFWPDQTKVTATQGKGDVVFRLT
SGLENFESNSILYKFQQKLYLVEALTHASYHENRATPSYQRLEFLGDALLDFLVTQHLYF
RHVNLSPGDLTDIRQALESIQHKKSPQKGEDHYSGNIPRNPIRLVLEKDKKAVFGKPKTL
DNGKIQRTLKVSWASEEFKGKGTNKKIAKKAAAKSAIKALEKKT
>maker-6776023_STYPI
MEEDAREKASESEAGGGADENLLARPYQVELLERAKERNTIVCLGTGTGKTFISVMLIKE
LAHEVRGKYMHKDDGRRTFFLVNTVLLASQQAKVIANHTDLRVKCYVGEMGVDGWEKSRW
ESEFNGNNVLVMTAQIFLNLLSAGFTTLSQVNLLIFDECHHARKNHPYAQIMEFFNRSKQ
LKPTEVTPLPKIMGLTASVVNGKVKLLRIESEIKQLECTMWSKCDTTCDEDVENFATKPK
EQVLSYSNEISEDLNKLIQHLSQALGRVMDFPGDCKVSNDEMIARGCAEWALEECTRTLH
ELGPWAAYIVAAYLINELETLQTSKEGKKICKSSATQLHQVQDVYCNFTGINGSGDDENC
SHIMPEYLTELLNIFGKFIGLESDPEKRLCAIVFVEKRYTALILSEQINQAAKLNHDLSF
VKSNFVTGHGTGGKVNFSSETEMNFKKQEEVLRKFRRHEFNVLIATSVVEEGLDVPKCNV
VCCFDFPKNFRSYVQSKGRARARDSNYYMLVPQELEGEKENDLEILREIEKILFKRCHDR
TQPSMRECIESLDTDALQPYVPVDGGGACVTMSNSISILSKYCSKLPGDRFTQPTPVYKI
EEIGKDSYRCKLTLPMNCQLREDIIGEPMRSKKYAKMAAAYKACIVLYRNGELDDNLRPK
RSLSDDESELEEEEEAAAAEPGGNESKPGSKKRKRQYTRKVPEVFVGSLVQDGQSQFLTT
ITIQVTKLIQPQEFVISERLFLNHTCTFGMLTNKRVPRMRSFKLYPNYGEVTVSLQCHTS
PLRRNNTKQDLDILREFHLFLFSHVLRSPVEFTPESMDGYFIVMLKDSGRSVDFDRMREE
LSRAHKQSRGNEPIVEDAVVTKNYVQCSQKYAVLKVRGDLTPLSPFPEKSQGNTYEEYFR
RKYGEMGRIADKNQPLVEAKELSGRVNFLVDRKVGTKHKRQVIYLVPELCNQIPIRASLL
SVSQILPSVLQRVTSLLLVADLKTMVAGEGDTTDESSLPPIGAEEGSHRPNQAGEVEDDG
LDLEDSDDAESEFPGLFSDIRSMFRYIPSSVNHPNSSLVLQAVTTTHSGDAFNLERLEML
GDSFLKLAVSLHLFCTYSDKDEGKLTRRKTNQISNLALYRAATKKSLGEYLQSTQLARDV
WCPTGCQFGDVPPNRTTEEREKRKKRKLTATEHASQKCITQMIGDKSIADSIEALIGAYL
ISDGYLGALRFMKFLGLKILPEVDPNDGIESLAKNSKSGCYARFWPDQTKVTATQGKGDV

VFRLTSGLENFESNSILYKFQQKLYLVEALTHASYHENRATPSYQRLEFLGDALLDFLVT
QHLYFRHVNLSPGDLTDIRQALVNNNIFATLAVEHHYHKYLKHMSPKWFQTMKDFIDR
>snap_masked-6652293_STYPI
MLTNKRVPWIRSFKLYPNYGEVTVSLQCHRSPLRTNNTKQDLDVLREFHLFLFRHVLRSP
VEFTPESMDGYFIVKLRASGRSVDFDNMKEELSRARKQSRGNVPIFVDALVTKNYVESPQ
KYAVLTIRGDLNPMSPFPEKSKGNTYEEYFRLKYREMGRIADKNQPLVEAKELSGRLNFL
VDRKVKTKNKRRVICLVPELCNQIPIRASLLSVSQILPSVLQRVTSLLLVADLKAMVAGV
RDPTDESILSPIEAEEGSREAPMEVEDNALDLVNDGDALFSDIRSMFRYSSSCAKHPDSF
LVLQAVTTTHSGDAFNLERLEMLGDSFLKLAVSLHLFCTYSDKDEGKLTRRKTNQISNLA
LYRAAAKKSLGEYLQSTQLARDVWCPTGCQFGDVPPNRTTGSSAMEVDSGDTVEAMEVDE
TCESGKKRKLTATENLRQKCNTQMIADKSVADCIEGLIGAYLISCGYLGALRFMKFLGLK
ILPEVDSDDDIDSLAENSKSGCYARFWPDQTNVTATQGKGDMVSRLTSGLENFENKAISY
NFQQKLYLVEALTHASYHENRVTPSYERLEFLGDALLDFLVTQHLYFRRVNLSPGQLTDI
RQAL
>Locus_14602_SYMB
MQPSTKKRSDDAVALRRFESLVRKLPRCLRAPSNVASDMQDISLHVHQLAVGEHPGKKSG
LALLLPEEVPVGAWSFSLLPPGHEQPILAKVKPATPGCPILLTSQQWEDLKVWTQLCLEI
ARCDRSCPLHQCFGQTMLTLWKDRSGKWLPAERTEVQSDKDLPERSFWLLAPLASADEES
AKISWSCLSWALAALAQFRESSSWLSMPRTWLGKFDPALPEMALDEAVVLGPMPTSSRGS
DVKGLCIDLEVKAGEEGAAIFTGYKFSRAAVNAAIDPEFAKVHRNLQQERLELELKAEDC
ELMPLTSGALRVLRCLPSLLWRLEFVALMQEMPLLCDGLLQTALPSALGEAMTHSKVLSL
PFQPSHPQWSKRFCYERMELMGDAVLKLMACTHAAAALPKASEGQLSSVAQWCETNKWLR
QVNEKTIQVGSYLLLESFRPKERLSKLRQGRVPQKVAADAVEAAIGAVFGCAAGAISEAA
EPLHPGASSLCLASGMDASWKIFRILVQQGPSTENESMDIKAAIAAPPCQNFQAAAVAGL
QTSLNLDAGTDAHAAEQVKNAFGYSFRNPKLLAALRASSTGARSVGFQRLEFLGDAVLLV
CVCCHLMQVCSDFDEGQLSQALQAFICNKYLSRKLIRRFGEVRSFASVFFPRQTSPQRVH
LPSQFDEVLASEVETDYVIGVRNVEAPGHKCVADGYEAMIAAVLLDTGGDLGETWAVFAK
DFEVPSRAELAELLRPRPRHEDHLHLDGSEKEEQTAKSSGLDAAVAPEFGQPEGEGRRLL

LDHCRRHGLKLRFEVKESGAVAEVRVRCVVGGRFLPEACGASMRQAQDLAAEAALWELTR
RSTQSDAKFPPPSLPESLRPGLTDFPQVDFEQSQRPKGIARQELQRYCNRNGLQHRFVES
EEEDVDHTVIHRMRVAAGDRIFPPATAKTKDAAQDMAAELALLELRSAEVTTQAPSSQGR
AREAPVETVFQDVSCLDTLPSAARQKGVDRQELQDYCSRSKLSLSFKEEAGPKPQPRTPG
LLRKRISEEMGPSHRPIFRVRAVVQGTFFPEASAETKNTAKDLAAAYALHCMRRDLV

5.3.2.2 Argonautes

>maker-6243026_STYPI
HTTRQTQGDGGPRPPKRPGYGTKGRPIPLRANFFRLNVSPGLSDLYHYDVEITPERCPKR
VKRDVVNEIINKYKKTTFQGHHPAFDGEKNLYSRIKLPNGKIPQDAVQALDIVVRQMPSV
YYTPIGRSFFPFDGQGRPLGEGCEVKFGFYSSIRHSEWKAMLVNIDGFYKDQPVVPDFLC
ETLDVKHHQLEDHRFELDRWRLEKAIRGVRIQTTHAASIKRKYTVWGVSEKSADRMQFDV
TDEGKGRTYKTTVGEYFKERYKLILRYPHLPCLQVGQKKDRYLPMEVCTIIPCHRKHLSE
QQTANMIRSTARPAPERQRDIQHWVQEMNRASSKYLKDEFQTSVNTEMVKVEGRVLPAPR
INLGPQDQPLVPRDGSWDMRNKSLHDGARIDKWALACFDRGCREEQLRNFSRQMASVSSR
QGLRMSEQPVVVAYGRGARDVESLFSKWVVDFPGLQLIMAVLPERDKEIYPELKRVGDNV
IGIPTQGVKSKNVYSCKPQLCANLALKINSKLGGINHVIDPREKSPVFREPVIIFGADVT
HPSPTENGIPSIAAVVASMDANATKYYARVRAQNHRNGKAAQEIINDLAAMVRELLIEFY
KANRKLKPNKIIFYRDGVSEGQFDQVLVHEVRAVQQACMDLEKDYRPRITFVVVQKRHHT
RLFCENQRDEVGKARNVPPGTTVDSGITHPYEFDFYLCSHYGIQGTSRPTHYHVLYDDNS
FTADDLQQLTYQLCHVYARCTRSVSMPAPAYYAHLVAFRARHHVTGNGSVDLEKSAKAIE
VNAKMKGAMYFT
>maker-6730085_STYPI
PEIDSKKTRHGLLKSHKEVLGPASAFDGATLFLPKQLESPTVLESERRTDGEKVTITVTF
TRVVPPDDCLQLLNIIFRRVMSRLHLTQVGRYYYDPHRPASIPQHKIELWPGYITSIQCY
EGGMMLLCDVSHRLLRTETCYDLMYNNKTYRIDDIAFDQNPTSTFTFHTGEQMSYVDYYS
KVYGIELQDLEQPLLIHRPKDKEQQKGRKLGLVCLVPELCNITGLTDAVRQDFRVMKDIA
AHTRVGPMQRQQAMLKFIDNINSCPEALQELTSWGVQLDQTMLQTEGRLLPFEKIILGST

SFISSPQADWGQQAVKEQVITPVPLRNWLVLYVNRDKSKAVEFVSMMNKVTPAMGIEVHQ
PNMLELRDDRTETYLRMIREHLNPQTQVVVVIFPTSRDDRYSAVKKLCCVESPVPSQVIN
AKTISQQNKLRSVTQKIALQINCKLGGELWALDIP
>maker-6734235_STYPI
MSSQGQGKNKRKGRGRGNAGRGRGRGHDLPPTPGTTDRVGKPLEKSGITIEEKPDYLNPK
DSSTADFDEKLAAGACGNDLLNDQKDTVRPKKSSTTASRATTTVEEHNQTGRSSVETPSK
SRKIKGRQRDIPASSSDDKGGGAPAADASSIKSVGATTPSSAAATSSTVAVNKEVENGTK
KLGSSLGSGGHTTHQTQGNGGPHPPKRPGHGTKGRPIALRANFFRLNISPELSDLYHYDV
KITPDKCPRSDKRDVVNKIIEEYKHTTFQGHHPAFDGAKNLYSRIKLPVPAELVVKLPGK
DGGKERNFKVKIQFAAAVSLLELNKFLSGKQNGKIPQDAVQALDIVVRQMPSLYYTPVGR
SFFPLDGRRSPLGAGCEVKFGFYSSIRHSEWKAMLVNIDVSAKGFDKEQAFVPDFLCETL
GVRAHNIEDRSFQPASWKLEKAIRGIRIQTTHAAPIKRKYTVWGFSGESAERMQFDVTDE
GTGRTYKTTIAEYFRDRYGLTLRYPHLPCLKVGQKKDRYLPMEVCTILPSPRKYLSEQQT
ANMIKSTARPAPERQSNIQHWAQRVTQASGKYLRDEFHTSISTEMVKVEGRVLPAPTINL
GPQDRPLVPLRGSWDMRDKSLHQGARINKWALACFDGRCHKDQLENFSKHMADVSSRQGL
TMSEQPVVVAYGRSARDVESLFSKWVVEIPELQLIMAVLPERDKQIYPELKRVGDNVIGI
PTQCVQSKHVHRINLQVCANIGLKINSKLGGINHDIDPGVKSPVFREPVIIFGAD
>maker-6778374_STYPI
HTTHQTQGNGGPHPPKRPGHGTKGRPIALRANFFRLNISPELSDLYHYDVKITPDKCPRS
DKRDVVNKIIEEYKHTTFQGHHPAFDGAKNLYSRIKLPVPAELVVKLPGKDGGKERNFKV
KIQFAAAVSLLELNKFLSGKQNGKIPQDAVQALDIVVRQMPSLYYTPVGRSFFPLDGRRS
PLGAGCEVKFGFYSSIRHSEWKAMLVNIDVSAKGFDKEQAFVPDFLCETLGVRAHNIEDR
SFQPASWKLEKAIRGIRIQTTHAAPIKRKYTVWGFSGESAERMQFDVTDEGTGRTYKTTI
AEYFRDRYGLTLRYPHLPCLKVGQKKDRYLPMEVCTILPSPRKYLSEQQTANMIKSTARP
APERQSNIQHWAQRVTQASGKYLRDEFHTSISTEMVKVEGRVLPAPTINLGPQDRPLVPL
RGSWDMRDKSLHQGARINKWALACFDGRCHKDQLENFSKHMADVSSRQGLTMSEQPVVVA
YGRSARDVESLFSKWVVEIPELQLIMAVLPERDKQIYPELKRVGDNVIGIPTQCVQSKHV
HRINLQVCANIGLKINSKLGGINHDIDPGVKSPVFREPVIIFGADVTHPSPTEXGIPSIA

AVVASMDINATKYCARVRAQNHENGKAAQEIINDLAAMVKELLIEFYKASGKLKPRKIIF
YRDGVSEGQFDQVLVHEVRAVQQACMDLEKDYRPRITFVVVQKRHHTRLFCENQRDEVGK
AGNVPPGTIVDSGITHPYEFDFYLCSHYGIQGTSRPTHYHVLYDDNLFTADSLQQLTYQL
CHVYARCTRSVSMPAPTYYAHLVAFRARYHVTGKEGSVDLEKSAKAIEVNAKMKGAMYFT
>Locus_1844_SYMB
PRTSCTQSWWSVMYRDRFLPQQSDGILQGNKQDARGTSGKKIKLSTNCFRISWKPQSFIH
HYVYELSVPDIQPDKERLVLEKAWESLKEKLQHFVVRCPGHIFSPNMASDFSVTADIGPN
ERVELKVCYSTKISGDQINTGAMGAASVVSKYMVDRFAEPLRIQKVGKRYYSNCPQAGKG
SHILISGSWVSSLVSSAGPLLQLDMIDRPVRKQSIVQILQASLEGADIFAHQTDRDVAAE
WIRCCVSATVVTSYNSRVYRIKQVHFDKDPSHTFMMYQRDQKEHMEISFAKYYEAFYHKT
IANKYQPLLEAYPEKETEKVFLLPELCSPTGITEDMRKEKQVLTDVLKQLKVQPQERLNS
IFSSVADMQRVQTPAVSTIQAWGCSLEKDPLEVQGRVLDPLQVCFKEKNYVIEEGNFTKY
LRHSIQVPIKIDHWLVIYLNEDEEVLKVWLKSIKDIAQCAFSMSVAEPHKVECNDQYTGL
SEILQENVTENTQLVMLITKENPKIYQLFKQKLCCMVPCISQVVKSETIRKRNGIAATLS
RLVLQMNAKFCGPLWHIAPVAAEDSEFEEFRKMPFMIVGVDVYQAYDGMKVLGLTATLDK
SYSQYFSESAILEPAWEPESWRASLSVNLQRLLRDAIACFARCNDKILPENIIVYRASVN
PEDWPDVAATELEAVKGVAGITGSCANAPYEPKITFITIAKRGNMRLFYRSEGDPACKNP
EPGTVVDDPAVCAGDVPNFYLISQAILKGSAIPAHFSVFANPANHSLEFLQNLTNRLCLM
YYTAAVAVRLPAPVVYAKKLASFVGSVIRKEPHPSLQRTLFYL
>maker-2878763_SYMB
MRDSLRGELGLRKLRVRAEYGQVKVKGKLVYGLTDKPANRYQFDCEEMGGKVDVAAYFKH
KHGLTLKHPDLPCIQLGNARNCIPMEYIYVMGGEHNLAVGKLRPEFQQEVTRRTAMQPSA
RRDQIMQALNNTQLGPSAALKPKGVDVAFEMLAVTGRLLDTPKLRDGSSAVMRDSTNYSN
NFKTLAPPKFNVSWGLWTFTTERSPSERDIQWFADQLTKRCYSKGMNFSDAKFVEWPEEA
FNSYFRCHKNREAFSPEMGNVIRKELNRVATAHKDLKLLVILLDNSEAITDHLYKLVKLI
TETEMGTFTTQFVNCKKGIEDAEKKLNNLMLKITPKLPEMPTGCRAAHNVALSEPHPLLR
KEEATMIMGADVTHKVAGISVAGVVGSTDSSYASYFHQIRGQSPYTLNTLKQRNRQSEER
IIDLTSMTYNILEKWRSMNKKLPDTIIYYRDGVSDGQFINVLKRELNLLDDAFKKISTTY

DPKLVIIVGQKRHQTRLWMQDGGLQDDKGKGKGKDKGKDDKGKGKGKGKDTAQVPPGTVA
SEGIAQPSHLNFFLVSQLGIQGTSVPCHYHILHLDKRLVKKGITVDDFETITFQLCHMYS
RADKSVGYATPAYMADHVCERGKHYLEANFGSSDVMSTHGSSVSEEKQDENLRKQIEERT
SWLNDRAAQSSRLNLNGLQLYC

>maker-3727646_SYMB

MSQEKLILTQGDGITAATKTQGDVAGKTFKVRVNLFRMNWTASIHQYGFALPETWIHSER
KLIEMGWADLKKNLSHFVVLLPGRIFSPIKVDKFPMKVSDASGGEVDVCHVAEISAQKIQ
DGLMGPASMVGQHLADQLAGQRRIQRVGKRFYKNDCQEVEGRHRVISGYSVNLAKLGSAG
PLLQLDVLHKPANTKSIVEVLRGSMEGTDVFQALPEIRAEWLRLCVSATVVTNYNFRVYR
IKQVHFDMNPSQTFQYHERGGGAKEYTYADYLSQYYQKSVTFNNQPILEAYPEKAKEKVF
LLPEFCCLVGVTDEMRKEKSSLSEALKQIKASPTERHHEIVRDAEEMEKQLPETLNAWGC
HLGQPIETLEVEAKQLEPLQVCFKTKQMSAIEEGSFSKSLRTGVQCQIMIDQWLLFYPEA
DEKAVDTWLASLRDVARLAFGCTLSQPGKVCYRDPFGELQSKIVEHQTPSTQLMMLLISG
KYERKVYNSFKLLACEQYPCISQVVRSETIGKSKTIPKQVIQQIHAKFGAPLWQILQDPE
DDGFRDFQRRPFMILGIDVYCTFEGARWLGICATLDKVFSQYHSMAAELENGSVQTWRTS
LSVELQRLFRDALSAFTDCNDGILPETIVVYRASVNQKEWPLVDAIEVQALLQVLNAAAK
LREGYVPKIVMLGIARKANMRIFKSDEGTIRNPHPGTVVDDPAVCPGPTPEFYMISQAIG
KGSAVPTHYSLLLNKAEMPLQLIENLTNRLCLMYYNVPSSVRVPAPVLYATKIAYFCGSV
IKKPPRPRLQRTLFFL

5.3.2.3 Pasha

>maker-6789121_STYPI

MSVDISVNELAVVVKAPTVLTVGENSELRTEIEKMENTMEKVSSSAGMDIPTYSKEESTD
EEMPFETERIPLKRKSEEDLLDHNHKRRMTEGDQADDWENVISTTGDEKQEGENPTDEQE
RKGNLPKLALPEGWIALNHRSGGIIYLHKPSRVCTWSRPYHIGGGSVRKHDVPLAAIPCL
HQKKGINQENTEETKSENGTSKSKNPAGSILNALNTQDVEKCVANTSVVKNTSEEQEETT
DSNKNTPPTLELLDIGELGSYLSTIWEFQTLTSEQERYAVMALPQTEDDVELPSSLECLS
YSVKGAENGKSPSKYCLLNAGGKTPVAILHEYCQRILKSKPVYLASESASSDSPFVAEVQ

IDGIKYGSGTGSSKKIARQIAAESTLEVLLPGMNQGICSSIEFKTVCGQNKSLSFIITCG
KHTATGPCKNKRNGKQLASQHILKKLHPHLEKWGALIRIYCDRPTGSIKKYKKDDSETAN
EKSGGNSSTNTDLLERLKSEMRKMYSENENANRMDAMDKPAEPVFTVTI

5.3.2.4 HEN1

>maker-6624208_STYPI
MTKKRESGKFARIWLGMGDTANDTTIIGHVARGDKPLLSTENEGSFKGPKFDPPVYRQRY
AAVCELVKERQAKKVLDFGCAEAKLVKTLISQDNLIHLEEVVGVDIDQQLEDSKFRIKPF
TADYLRPRSHPFKVSLYQGSIAEADERIEHLEPDILDLMPKAVFGQLSPKVVMVTTPNVE
FNVLFPDLKGFRHYDHKSNTQALKYNYTVRFDGIAHGPKGTEHLGCCSQMAVFERMSSFS
ASRKESGIGQPYNLIAEVQFPFKKDTRTEEEKILQEVEYILWMLSSQEEVHDSEESDVDP
LSSDKDETSCHYDQNHLMHDEDCAKGGNDEAQRHVYLLKELLGFRSLQKFCNDIEKLRSV
LKGSSRFCLTEDERGVLWQHQSSSQWSSEESDTGWDVCGDEQEVYESKRPDSAWTEPENW
DAADEITTDANVKQTEANTEVCWDSANGCNNKCWNGEDWSKLEEYGADDEFQEENSFNFH
NSYTVAVPLESDGSGSDDIDGNEIGDLS
>maker-6785799_STYPI
MVKEYCLMAKNSARKEWKLSSGDWTTQTEGPAMTKKRESGKFARIWLGMGDTANDTTIIG
HVARGDKPLLSTENEGSFKGPKFDPPVYRQRYAAVCELVKERQAKKVLDFGCAEAKLVKT
LISQDNLIHLEEVVGVDIDQQLEDSKFRIKPFTADYLRPRSHPFKVSLYQGSIAEADERI
EHLEPDILDLMPKAVFGQLSPKVVMVTTPNVEFNVLFPDLKGFRHYDHKSNTQALKYNYT
VRFDGIAHGPKGTEHLGCCSQMAVFERMSSFSASRKESGIGQPYNLIAEVQFPFKKDTRT
EEEKILQEVEYILWMLSSQEEVHDSEESDVDPLSSDKDETSCHYDQNHLMHDEDCAKGGN
DEAQRHVYLLKELLGFRSLQKFCNDIEKLRSVLKGSSRFCLTEDERGVLWQHQSSSQWSS
EESDTGWDVCGDEQEVYESKRPDSAWTEPENWDAADEITTDANVKQTEANTEVCWDSANG
CNNKCWNGEDWSKLEEYGADDEFQEENSFNFHNSYTVAVPLESDGSGSDDIDGNEIVLRY
LISFLMCCKD
>Locus_19351_SYMB
VISEPGTAMGRDLSEKIELLGIAVADRLAQVSGDHSDLFKRFIDAQDRVPISVLLMDKGL

RRTALKLAGEGGRSLARAFEHLQKAAEANEDLAVTHLPRTMVRQAKQKEMPPAVAEAMRL
VEAEEAQARVLAEAASSAQPALSIPEDSPKAAPVALQPAQRNVKGELGHFFAKKMGRGAA
KGDINFVAQRVSDHPATFRAQCEIWGRFFESKDAHTTKKAAEHDAALTALETLLAEAGMK
DSGAAAVDWKGQLLSFFQPQGQREQVQFEIQRVGQEESEEKIFQAKVALPSGQVFLSEPQ
SRTVHGCSAKTRRAAEQDVARKAFLALQGKADSEASPSPSPTPSAAFQPPPRPTHREVQI
LWLPAAAHQMPFLRRVVVESDDFLATAESLLKCSNTVCYPLSWQRDRPKGLSLYEAIDAS
EVVNDRATAIAGVEVFGDSFMMDASAHSSGAELADITLPRYEFLKAERAGFLREPVPALN
HSRGMLVCGHPKFPESFSAETWVGLNPRSLLNDVVRRKGLAQPRTDVILCKDDLPKEAAK
GQSWFQGVISLPELGLQVHGAMATNKSEAQQLAAVKMLQRLQEEAEAAGAVWQHPLCKSA
MAPAVRGVPRPSRKPLPPGSKVICSYQITLVAESDQDGPSIELPLEGQERLTAFVGGNIL
HPVLEDLLQRLTPAGEAVTVEAPCEYEGNPCKFVMSAQLLSITHPMEPPPEPAKTIKFDP
PLWKQRQRYIVQALQARKVSSVLDLGCGDGQLLEALVAVGLQRVLGLDLSESRIKAASRR
LAEKGGQEQHVEVMQADFVNVPAEGEPRWISFAAGVQAIVLCEVLEHLPEVAMPRLPGAL
FGALQPEVVVVTSPNADFGESSDSEDEDAGPAEPRQFRHADHEREWTRAEFRTWAEAAAE
KHGYWISEFDGVGYLPDQESQGPCTQLAVFERKDLEEAAEEVEEAAAFSGSETQVDLSSI
SAKALRSQELDSEGSGLWKHVLREGAGEPPPQHSRVNLHYSTYLIDGRRVDTSRGGRRSM
PCAFQLGRQDQGVEAWQLAAQTMSCGEVAWLQCPARYAYGSQGAPNIPADADVWFCLELT
STKAPGTVRFLSRVADALDEAEKHMEVGRADIKRQAFAQARQAFRRAVAAVPDKLLLKQS
DAIITRFAKMERASLLNQAHCCLKLGDSVPDVKSEQQAHFREALQACRSLFDRHGPRDVQ
AEDGCLPESMVQVADAICKASQLEEVPWMAKAHFRRGLAKEKLRYLTDAVADYEAALALE
PTDSVISKQLQLVKARKQKAELKPTQMFAGILQRERAEREREEAAADLAARKQRRKERLQ
REAEQREAQKPDQASET

5.3.3 List of condition-specific, overexpressed short reads

5.3.3.1 4C (Extreme cold stress)

AAAGGATTCTAATAATGGTAGTTTTGATACTC

AAAGGATTCTAATTTTTGATACTCATTCTGTT

AAATAAGCTTTGGCTCTAACGCATACCAGCCT

AACACAAGCCCTCAAGCAAGATCCTTTCAGAT

AACAGTGGCATGCTCAGCCTCGCTGTGCCGGA

AACTCATAAAGTCTGAGCTTGCAAGCCGTGAC

AAGCTGCGTGGGTTCAAATCCCACAGCTGTCA

AAGGACAGTTTGGCCGAGTGGTCTAAGGCGCT

AAGGATGACACGCACAAATCGAGAAGTGTACC

AAGGCTCTTAACCTTGTGGTCGTGGGTTCGGC

AAGGCTTTGCTTGTGATCGCTCTGCGATGACC

AAGGGCGTGGGTTCGTACCCCACAGCTGTCAC

AATGAAAACTTTAAGTGGGCCGTTTTTGCTTG

AATGAAAGCTTAGAGAAGATTGGATCTGAAGA

AATGGACAAGGCTTTGCTCTCTCGAAGCCCTT

AATGGACAAGGCTTTGCTTGTGATCGCTCTGC

AATTTCATGCAGTGGGAGCGATAGATCTTTTG

ACACGTGGCTCCAGACCGTCGCTCAGTTGTGG

ACAGAGAGACCTTGAGCCATTGTCTGAGATGG

ACCAAAACCCAACTTCGCAGAAGG

ACCAGACCTAGTAGCCTAATGGATAAGGCGTT

ACCTGGGTACATAGCTCAGTGGTTAGAGCGGC

ACCTTTGGCCTTGAGTTGTCTGACTCCCTGAG

ACGATGATGCTAAGAGAAGACGTCTGAGGTCT

ACGCAGAAGAACTCAACGCCCTTCGTCTGTGG

ACGCGGGAGAGTAGGTCGCTGCCAGGTCTTGG

ACTACTGTCCCCAAGTCCTGCCCT

ACTAGCTTGTGAGCTGTTGTGAGCTGTTGCCA

ACTCAAGTTCCACTAGCGCAAGCTGCGTGGGT

ACTCAATTCTAGCTCAAATCGGACTCCAGCAC

ACTCACACAATTTTTGGGGAGCTCTGAGCTCC

ACTCGCTGTGGCACCCTGATTAAAGCTCCGTG

ACTCGGTGAGAATAGCTACAGTGACCGCCCCA

AGAAGCATCAGTGGTCTAGTGGTAGAATATGG

AGACAACATTCTCATGAAAGCAGCACTGAATT

AGACATGATACTGATCTCGGGCACACTGACAT

AGCTGCCGGAAGACAGAGAGACCTTGAGCCAT

AGGACTTGGAACAGCCGGGATAGGCAATGCCT

AGGTTGCCGTGAATGGGACGCATCAATGTGAC

AGTACGGCTGCTTGATGACGAAGCACCAGTTG

AGTAGTCTATGCAAACCTGGCAGGCAAAGATT

AGTCATCCATAGCAGAATTTGTTGAGAGGCAA

AGTTTTGATACTCATTCTGTCAAGGTATTGGT

ATAGAGTAGTCTGTGCAAACCTGG

ATAGCAGAATTTGTTGAGAGGCAACTAGCTTG

ATATGATTGTGTGAGCAACCTGCTATAATCTT

ATCAATGTGACTCACACAATTTTTGGGGAGCT

ATCACAACCCACATGAGATTTCTCTGACATGG

ATCACAGCGTTCACTTCATAGCGCACCAATCT

ATCACGTTTGGTCGTCTCACCAAGGAGAGCTC

ATCCGGTGAATTATTCGGACTGACGCAGTTGG

ATCCGTGATGTTAAGCGTGGACAGATTGTGAG

ATCTTTTGTAGACGACTTAATCGAATCCAGGT

ATGCAAAAGACCGTAGTTTTCAACCTGAGATG

ATGGGCAGACACGCTGTGGCACCCTGATTACA

ATGGGTTAGTCGATCCTAAGAGAGAGGGATTG

ATGTGACGACTCTAACGATATGACGAACGATG

ATTTCATCCGACAGGCCTGTCTCATTCCTTGG

ATTTGTTCACCAAAAGACTTTGATCCTGATGT

ATTTGTTGAGAGGCAACTAGCTTGTGAGCTGT

ATTTTTCAGCCATGTGCTCTTCGCAGTTCTCC

CAAAGGATGACAGAACATGCATGTATTAGCTG

CAAATCGCTGGTGTCGTCTTGTGGCTGTTTGG

CAAATGCTGAACTTGTGGATCGAGAGCTCTGA

CAAATGCTGAAGCTTCGGATCAAGATGTCTGA

CAAATTATGGTGGTTCGTGTCCACCCTGGGTC

CAAGGTCATTCTTGCAATTTTTACGTGGGCTT

CAATGGGGTGAATGACAATTTCATGCAGTGGG

CACAAGGACATGCTTCACTGCTTGATGAAAGA

CACAAGGATGAGAAAAAGTGTTGGCACCAGAG

CACACTGTAAGGCTACTCACCTCCAGCAGGTG

CACAGGATGACTCAGAGATAAATTCGGCGGAG

CACAGGATGAGATCACAATTGATAGGAAGAGT

CACAGTGGTTACTTGATTGAACTTCTGGGATT

CACCGAGACTGGAGGGCGCCGGGTGGCGGCAC

CACCGTTCTCATTGCAACATGCCAACAAATAA

CACGCTTAATGAAAGCTTAGAGAAGATTGGAT

CAGAAGCCAGCAAATTATGGTGGTTCGTGTCC

CAGTACCACCGCCGGTGCGTCGACGTCGGATG

CAGTCCGCAAGGGCGTGGGTTCGTACCCCACA

CATAGCACCAAGCCAGTGGCAGCACACCCTTG

CATAGCGCACCAATCTTTCGCCTTTTACTAAA

CATAGGATGTCCGGTCATGCAGTACATGTTGG

CATCATGTGAGTTTTTGATTTAACGTTGCCTC

CATGCAGACTTGAACAAGCAAACTATGGTGGT

CATGGGTTCGTACCCCATAGCTGACACCATGG

CATGGTGGTTCGGGTCCACCCTGGGTCGCCAT

CCATATCTCACCGAATGCACCGGATCTCTTTC

CCATCGATGTGACCCGGGTTTCCCGGCTGATT

CCATTCGGATTGCAGCTGCAAGTTCTGACACA

CCCAAAGGTCCCTGGATCGAAACCAGGCTCCG

CCCAAGACACGTACTGACTTCAAGCTGAGCCA

CCCACAGCAAAGTATGGTGGTTCGTGTCCACC

CCCGTGGCTAGCTTGTGAGCTGTTGCCA

CCCGTGGGCATCAGTGGTCTAGTGGTAGAATA

CCGAATGCACCGGATCTCTTTCGACCTCCGAA

CCGATGGAGTCATCCATAGCAGAATTTGTTGA

CCGCACGTCTGAGAACTCAAGTGTGACAATAC

CCTAGTTAATTATATGATTGTGTGAGCAACCT

CCTCAGTTCGCTCTAGAATTACCCTGAGCTGG

CCTTCCATGCTTGTGCCCCGGGTTCGGCTCCC

CCTTCCGGCAGAAGCCAGCAAATTATGGTGGT

CCTTCGATGTCGGCTCTTCCTATCATTGATGG

CGAAGCAACTTTAACTATACGCTCTGAGATGG

CGACCGATGCCCGACCACTGGTGC

CGCAGGATGAGTTCTGATGAAGTCACATACAG

CGCATTGTCTTCGACCTATTCTCAAACT

CGCCTTACCCGATGGCTCGTCATCTGTGTACT

CGCTAAAGGCAATGGGGTGAATGACAATTTCA

CGCTCTGCGATGACCCACAGCAAAGTATGGTG

CGCTGTTTGCTAGTTGAAAACTACTCCAATTC

CGGCTCTCACCCGAGCGACCCGGGTTCGTGTC

CGTAAGGGCGTGGGTTCGGACCCCACATGAAT

CGTTCCCGTGGCGTAACTTCGGGAAAAGGATT

CTAATTATATATCCACCAAAACATCACTCCCC

CTATCGCCGCTAAAGGCAATGGGGTGAATGAC

CTCAAGGTACAAGTCGGGCCCTTGCTTCACAG

CTCACCGAATGCACCGGATCTCTTTGGACTGG

CTCATGATTACGATCGAGGCATTTGTCAGAGG

CTCATTCTGTCAAGGTATTGGTTCAAGTCCAT

CTCCATGTGGAGAGCTCACCTCCC

CTCCGAAGTTAAGCGGTGGAGAGCCCGCCTAG

CTGACTCCCTGAGCTCACACTGTAAGGCTACT

CTGAGCTCACACTGTAAGGCTACTCACCTCCA

CTGAGGATGATCGCTGATAAACCTTGGAGATC

CTGGACATGGCTTCAGTCCATGGGGCATGGGG

CTGGCCTGGGGGACCGGCTGGGAAGCGCTTAG

CTGTTGAGATTAGTAAGAAAAAAGAAAAAATG

CTGTTGTCCATTCGGATTGCAGCTGCAAGTTC

CTTAATTCTGATGAGCGTGTACTGAGACATGG

CTTCGTTAACAACTCATACGCTATTGGATCTG

CTTGAACAAGCAAACTATGGTGGTTCGGGTCC

CTTGCAATTTTTACGTGGGCTTCGGCCCATGG

CTTTCGACCTCCGAAGTTAAGCGGTGGAGAGC

CTTTGATTTCTGCCAGCATGGATTTTCTGAGG

CTTTGCGTTTGCCCTTCCGGCAGAAGCCAGCA

CTTTGGCTCTAACGCATACCAGCCTCTATCGC

GAAACAGGGTGGCCTGATAGTACTTCTGCTGG

GAAAGGATGATACGTCCTTCGGACACTGTGAC

GAACTCAACCGAGATCTGTACTGACGCCCTGG

GAAGCCATAGCTGCCGGAAGACAGAGAGACCT

GAATGATTAGAGGAATCGGGGACGCGTTGTCT

GACACTTTGGCCGAGCGGTTAAGGCTTCGGCT

GACAGGAAGAGACTTGAGATACGAACCGTTTC

GACCCAGTAGCCTAATGGACAAGGCTTTGCTT

GAGACGCAGAAGTCCCAAAGTGTCGGATTTGG

GAGCAGTTCACTGCACTCTGAGGCCTGTAACT

GAGGCAGAGAAGCTGCTGTGAATAAATTGGCA

GAGGCCTGTAACGCCTTAGGAGTCTCCGATGG

GAGGCGCCAGTCCGCAAGGGCGTGGGTTCGTA

GAGGCTTTCATTTTGGAAACTTAAATGTTCTG

GAGGTACTCTTTGCACCTTTGGCCTTGAGTTG

GAGGTCGAACATCAAGAGTAGTGGGCTGAGTT

GATATGATGACTCTATGGCGCTCCGACACTTT

GCAAACTCTGGAATCTCGTTAAGCTGATGTGG

GCAATCCGTGTTGACCACATCAATCTTTGAGG

GCACAGAGATAAGTCGGGCTTTCACAGAGTGG

GCAGCTCGCTCTCATAGGGTCGAGCAGTTCAC

GCATCAGTGGTCTAGTGGTAGAATACATCTGG

GCCAGGGTGAGTCAGGCTGGTGCCTGGACATG

GCCAGGTAGTTTGTGACGTGCGAGGGTCAGCT

GCCGCGACAACCCAGGTTCGGCTCCTGGTGGG

GCCGTGCTGTCGTTGACACAGTGAGAAACTGA

GCCTCTGGGCCTGTTGTTGAACATTTCTCTGA

GCCTGTCATGTGGAGAGCTCACCT

GCCTGTGACGAGAGCTTTCTGAGCCTGACATG

GCGATAGATCTTTTGTAGACGACTTAATCGAA

GCGGACATAGTTTAGTGGTAAAACCTCAGCCT

GCGGAGATCCAGCAGCTCGCTCTCATAGGGTC

GCTCTCATAGGGTCGAGCAGTTCACTGCACTC

GCTGACGGCCATATCTCACCGAATGCACCGGA

GCTGCTGCATGATTGCCTCAGATCTCTGAAGT

GCTGGGATCATATCATCCTGTCACCTCTGGGG

GCTTAGCGGAGGTGCGCTGTTTGCTAGTTGAA

GGAGAGAGTGGGCTTGAATGGATGGCAGTTGG

GGCATGATGATCTTAAAAGCTTTCCCCGCTGA

GGCGAGGACGATGCAACAAGAACAGAGCACTG

GGGGACATGCAGGGGTTCGTTTCCCCCTCTCT

GGGGACCGGCTGGGAAGCGCTTAGGGTGCTGT

GGGGATTTAGCTCAGTGGTAGAGCGCGCTTGG

GGTGCGAGAGGCCTGGGTTCATTTCCCAGAAC

GTAATCGAAGTCAGTACGAGAGGAACACTTGG

GTACAAATCGGGCCCTTGCTTCACAGCAGTGG

GTATCCTTTGTGATACGATCCACTGAGATTCA

GTGAATGACAATTTCATGCAGTGGGAGCGATA

GTGACAATGTGATATCAGAAGGCGAATCTGAG

GTGACCTCTGGTGGCATGAATGTGTGTCACTC

GTGCTTCACGAGTTGGGCGCTCCTGTTTCGGC

GTGGAGAGCCCGCCTAGTACTGGCCTGGGGGA

GTGGCCGCATGTGGGACCTCCTTAGATGTCTG

GTGTCTTCATGAGCTCGCTCATAAGCCT

GTTAAGCGGTGGAGAGCCCGCCTAGTACTGGC

GTTAAGTCGCATCGAGCCCGTCTAGTACTATC

GTTACATGGGCAACTTCTTAGAGGGACTTTTG

GTTACTAACCGGCTTAGAGACCAGTCCAACCG

GTTATAGCTGGAGAACCTGGTGTTGAAAAAAA

GTTGAAAACTACTCCAATTCCCGTCAAGTTTG

GTTGATGGCGAGTGGGGCAAACCATGGGAGCA

GTTGCTGTGGCGATACTCTTGTGCTTGAATTC

TAACCAGGTATCGTCTTTTGTGATTTGAGGCT

TAAGACTTGGAGATGAAGACAGAAGACTCATG

TACCTGGCTAGTGTTGGGCGGAGATCCAGCAG

TACGTGAGGATGGCCCTTGGCTCATGCTGTGG

TAGACGACTTAATCGAATCCAGGTATTGTAAG

TAGATCAGATGGTCCCTGGTTCGTTTCCGGGT

TAGGGTCGAGCAGTTCACTGCACTCTGAGGCC

TCACTTCATAGCGCACCAATCTTTCGCCTTGG

TCCAGGATGAGTCTTCAAGTTTCCATTCGGCA

TCCGATGATGGACCATTAGCGCACATCCTGAG

TCCTTGATGACTTGCCGGCTCACCGTCAGAGG

TCGATTCCCGGCTGATGCACCAAAAAAAAAAA

TCGGCATCAACAGAGTACAAAGAACCCTTATT

TCGGCCACGCAGGAGAGAGCCCTGGGAAGTGG

TCTAATAATGGTAGTTTTGATACTCATTCTGT

TCTCGGATGACAACCTAACTTCTGACACCTCT

TCTTTGCACCTTTGGCCTTGAGTTGTCTGACT

TGCGTGATGATCGCTTTTGTAAGATGAGATGC

TGTAAAATGATAACGCAGGTGTCCTAAGGTAA

TGTCAAGGTATTGGTTCAAGTCCATTATCCTT

TGTTAACTTCCAATTGATACTCCTGAGTCTTG

TTAACGAACGAGACCTTAACCTGCTAAATTGG

TTCACAGCCTGTGAAGCATCGCCCGAGACTGA

TTCACTGCACTCTGAGGCCTGTAACGCCTTAG

TTCGTGATGATTGTCACTCTCTTAGGACACCT

TTCTTTTTAACCGCCTTTTACCCTGGAATCAT

TTGAAGAGCCTTTTGTTCCGACACTTTGTCTG

TTGATGATGAAGATTGATCCTCTCACAAGGAC

TTGCCATCGATGTGACCCGGGTTCGTTTCCCG

TTGTAAACTGGAGCCCTGGGTTCATTCCCCAG

TTGTAATATCTGCTTCGAATTCAGGAATAGGC

TTGTGAGCTGTTGCCAACTGAATGTGTTTTTT

TTTACTTATTCCATGAGACTGAAACTTGCTGG

TTTATATGACTCCGCCAGCACCTC

TTTATATTTAAATTAGACTCACTTCCCTTGGT

TTTCATATCTTGGTCTAACCGCTCACTAGAAA

TTTGCAATCTTGCCTGAAATGGCC

TTTGCGTGCGGGAGGCCCTGGGTTCATTCCCC

TTTTGTTCGCCTCGAATTGTAGTTTCGAGATG

5.3.3.2 16C (Cold stress)

AAAGAGGAGTTCGATTCTCCTAGGGGCTGCCA

ATTGGTTGACCAGAGGCCCCGGACAAATGTTC

CAAGGTCTGTCGTCAGCTCTTCATACCAAAGA

CAGGCGACTGTTTATCAAAAACATAGGGCTGG

CATCGCAGGGTAGCTACGTCTGGAAGAGATAA

CCAGAGGCCCCGGACAAATGTTCGGGGCCTCG

CGGAGTGTAGCGCAGCCTGGTAGCGCACCTTG

CGGGCCGTGGCGCAGCCTGGTAGCGCGTTTGA

GACACGTGAAATCCTGTGTGAATCTGCCAGAA

GCGGAAACAGGGGTTCGAGTCCCCTTGGGGGT

GCGGGTGTAGCTCAGTTGGTTAGAGTGCCGTG

GGAGGGTATGCGAATTGGTGAGCGTCCTGATT

GGGGCCGTAGCTCAGATGGGAGAGCGCTGTGG

TAACATCTGGGAGAGTATGTCGTTGCCAATCT

TTCGAGATGAGGACGTTGATAGGCTATAGGTG

TTTCGCCAGGCCTGCAAGAGGGGCGCGCCAAC

5.3.3.3 20g (Hypo-osmotic stress)

AAAGATTGAAATCTTCATGAACGACTCAAAAG

AAAGGTCCGGAGTTCGACTCTCTGTCGCCTGG

AACGGGCTCAAAAGTAGGATTAATAGGATCCA

AACGTTGCCTCTTTTGGCAGCTGATCTGTGCG

AAGAATCCTGGTAATCGAGAGAGATGTCGTAA

AAGGAAAAGTCGTAACAAGGTTTCCGTAGGAA

AAGGTCGGTGGTTCGACTCCACCAGTGGGCAC

AATAAGATATGAAGAGAGGATGGATTCTAAAA

AATAATTGGTAGCTTATTTCTTTGATATTGGC

AATCTTTATTTGGATAATGAATAGAAATCTAC

AATGTTCTGGTGGTGATTTTGAACAATAGTGG

AATTAAAGATGAGTACAGCTGAACGTCCATTG

AATTAGTTCTCAATGTAAAAGATTAATAATCT

AATTGTCGAATTTCAATCACTGTCCTTCACAC

AATTTTGCACATACTGCCCGTCAAGCAAAAAG

ACACTCAAGTACGGCATCTCTGATACGACTCG

ACATAAGACCCTGAGAACTTGGATGTAGAACT

ACCCTGAGAACTTGGATGTAGAACTCAAGCCT

AGAAACCGTACCATTGTGAGTCTTTCTACGTG

AGACGGTGGTTCGACTCCGCCCTGTGGTACCA

AGAGAGATCAATGGTACGAAAAAGAGAAATGG

AGCCTTACACGCTGAAGGTCGCCAGTTCGTTC

AGCTGGTTAGAATACTCGGCCCTCACCCGAGC

AGGATTGGCTCTGAGGGTTGGGCACATGAATT

AGTAGGATTAATAGGATCCATCTTTTCAAAAG

AGTGTCGGATTGTTCACCCGCCAATAGGGTGG

ATAAAGTCATGACTGAAGCTCTAAATGTCGAG

ATAAGAGATTTTTATTGGAAGAATCCTCAATC

ATACTTGTAAAATGTGGGCCTCATTAACCTTG

ATAGAGCAAAGTAGGTAAGGGAAGTCGGCTGG

ATCAGTACAAAGGATATGCTGGTGAGCTGGGA

ATCTTTGACAATTTGGTGAGTGACCTCCTTGG

ATGAGTACAGCTGAACGTCCATTGATTGATAT

ATGAGTAGAATTTAAACCCCTTTATGAGTTGG

ATGGAATGCTCTTCGGGGCGTTCCCTGTGCAG

ATGGCTGCTCTCAGCTAATGTTTTGTAATATC

ATTAATGATTATATCCTAGATCCAGAGGTGAA

ATTCTATTATCTGGCTTATTAATTCTGGCAGC

ATTTATGGCTTTATAGTAATGTTTGTAGCTGT

ATTTGGAATCTTTGGTTAGTTACTTAAAGGTG

CACGGGTTCGAATCCTGTTGCCGGCACCATGG

CACTGTCCTTCACACTATAAAGTGTTTAAAGC

CAGAAGAACGAAACCATTAATTTAAACTAAAG

CAGGTCATTTTGCACTAGAATTGGAGTTATTT

CATTGCAACATGCCAACAAATAAGCTTTGGCT

CCATTTATCGAAAAGATACCTCTTTGTAGATA

CCGCCTAGTACTGGCCTGGGGGACCGGCTGGG

CCGTCAAGTTTGATCGTCAAGGTCATTCTTGC

CGAGTGGTCTAAGGCGATGGCTTCAGGCGCCA

CGCCTCGCGGATCGGCACTGCAACCGGAAGGG

CTAGTGATGATTCCTTGATAAAGTCGGACTAG

CTCGATCGGCTTGATCACTCTCATTAGTCTGG

CTGCGGCTTAATTTGACCCAACACGGGGATGG

CTGGGGCGGCACATCTGAAATGATAACGCAGG

CTTAGTAGGTTTCCCGGGATTGCAAACTCCAA

CTTAGTGAGCTTAAAATTAATTTACCAGCACC

CTTTATAGTAATGTTTGTAGCTGTAGTAACAC

GAATGGGACGCATCAATGTGACTCACACAATT

GAGGTCCCGAGTTCGATCCTCGGTCGAAGTAC

GCCGGATTAGCTTAGTGGTAGAGCATCTGTCT

GCGAGGACGATAATGAGCTCCCTCTACTGACG

GGAGTGATGTCCTTGATGCATCACAGACCTGA

GGTGTTAAAATGCCGATTGTTAATAGTTAAAT

GTAATGTTTGTAGCTGTAGTAACACATGTTAT

GTCATGGTGAAATATTCATCAGTACAAAGGAT

GTCTCTAGGAGTTAATCTGTTGAGTCGTGAAG

GTGCAAATCGTTCGTCATACTTGGGTA

GTGGGTGTAGCACAGTGGTAGATGCTGAGGAC

TAAAATGTGGGCCTCATTAACCTTGCTGCAAA

TAACACATGTTAGGTAGGAGTGGACTCATAGT

TAGAGTTAAAAGTGCTTGAAATCGCTGAACTT

TAGCTGTAGTAACACATGTTAGGTAGGAGTGG

TAGGAGGTGTAGCATAAGTGGGAGCTTCGGTG

TAGGTAGGAGTGGACTCATAGTATAATTGTGA

TAGTTGGTAGAGCACCCGCTTGGTAAGCGGGT

TCAGACGATAGAAAGGGAAGCATCCATAAGAG

TCCGGCGTAGTGTAATGGTTATCACACCAGGT

TCGGGAGTTCGACTCTCTTCATCGGCACCATG

TGCATGATGCGTCTACGCAAGAAACTCATAAA

TGCTGAGAGTATCCTTTGTGATACGATCCACT

TGGTGGTGATTTTGAACAATAGTGGAATAGAT

TGTAAGTGCTGAGAGTATCCTTTGTGATATGG

TTAAATGTTATACGAGCCAATGTTAATATCTG

TTAAGGAGGAGGCTTTCATTTTGGAAACTTAA

TTAAGTGTAAATCGTGCTTTAATTTTTTTTGG

TTCCTGCTGGTGTATCGGTAACATTGGACGTT

TTGAGTAAGAGCATGTGTGTTAGGACCCGTGG

TTGCGGGTTCGACCCCCGCATCGGGCTCCATG

TTGGATAATGAATAGAAATCTACTAAGCGGTT

TTGTAAATGCTAATTCCTAGTTAATTATATGA

TTTCAATCACTGTCCTTCACACTATAAAGTGT

TTTCCGGTAGTGTAATGGTATCACTCGTGCCT

TTTTGATTTAACGTTGCCTCTTTTGGCAGCTG

5.3.3.4 36C (Extreme heat stress)

ACCAATCAATTGGGAGAAGTTTGA

CACCAGCAACCGACCAATCAATTG

CACCCTGTAGAACCGAGCTTTGGTTT

CAGCAGCCGCGGTAATTCCAGCTC

CATCCAACCAATAGTTAGCAGTTAAATGTTAT

CATGGAAGAATCAGGAGTAGAAGTATACTTGG

CCAGAAGGACGCCCCATATGGCTTGGGCGGGC

CCCGAGAGACCCATTAGAGGCGAAGTGAGAGC

CCCGTGGCAACACGGGGAAACTTACCAGGTCC

CCCTCAAGCAAGATCCTTTCAGATATAGCTGA

CTAAGGAGATGGACTCGAAATCCATTGGGCTC

CTGGCTGGAGCATTGCCATTGGTGCCTTGGGC

CTTAGGAGTCTCCGATGGAGTCATCCATAGCA

CTTGTCTCAAAGATTAAGCCATGC

GCACTCAACGTCTTTCGCTCCTATCTGAGCCT

GTCACATCTATTTTTGCTGGGGAA

GTCCTTATCGTCTAGTGGTAGGACTTCGCGTT

TACCCTGTAGATCCGAATATTGT

TATCAAGCTACTAAGGGCTTACGGTGAATTCT

TCTGGTAATCGAGAGAGATGTCGAAAAAAAAA

TGGACGGAGAACTGATAAGGGCTGGAATTTCT

TTCGGTGATCTTGAAATACCACTA

TTTGAAATCCATTGGGCTTTGCCCGCAGGGGT

5.3.3.5 DC (Cells harvested at midnight)

AAAGGATATGCTGGTGAGCTGGGACCGCATGG

AAATCTCTGCAGAATAACGAGCATACTGATGG

AAGGACCAGAAAGATTAAGCTAAATCTGAATT

AATCGGTTCCAGCGACTGTTTATCAAAAACAC

AGATCTCGTATGCCGTCTTCTGCTTGAAAAAA

AGCAGACCGTCCGTACTAAAACTGACACAGGT

AGGGTTGAAGCAGACCGTCCGTACTAAAACTG

AGGTAGCGAAATTCCTTGTCGGGTAAGTTTGG

AGTACTAATTGCTCGTGAGGCTTGACCCTTGG

AGTGGCAAAGGATGTAGGACTCCC

ATGCATGAAGCTTACCGGTACTAATAGCTCGT

ATTACAAGTAGGACGGGCGTCAGAGGCAGCTG

ATTTATGGTAGGAGAGCGCCGCTGAGCAGATA

CAAATAAGATTCATCGTCCTAAGCATACATCA

CAACTAGAGATTGGAGGTCGTTACTTATATGG

CAAGACGTGGGAGAGTCGGTCGCCGCCAGGCC

CACCAGGAGTGGGAGGGCATGAAGCAGACCAC

CCACAAAGTGCACCCCAGTTCGGATTGCAGGC

CCCCGGTTCGATTCCGGGCGAGGCCTCCATGG

CCCGTGTGAAAGTAGGTCATCGTCAGGCTTGG

CGAGAGATTGGCCCGCGTTGGATTAGCTATGG

CGCAGAAGGGTTGTATTTATTAGA

CGTCCAAACGTTATTCGGAATTAT

CGTTGGTTCAACTCCAACCGTCTGTACCATGG

CTCACCAGTGAGTAGGAGGGCGCA

CTCATCAGGAATGATGGGCTTGATTTATCCTG

CTTCGTCTGCCACGCTCCGGCCAGTGCAAGTG

GAAAGGTCTAGGGGTGTAGCTCAGCTGGTAGA

GACTTAGGAGGTGTAGCATAAGTGGGAATTCT

GAGAGGATTTAGAAGGACGTCCCT

GAGGTTTCCTCGGCGGGGTTGTAGGGCCCTGG

GAGTGGAAGGACCTTGTGAAGTTCTGAGCTGG

GGAGCAGCAGACACGTTGTATTTGTAAAAAGG

GGCCCGTTCGTCTAGTGGTTAGGACTCCAGGT

GGCCGATCCGTCGACCTGGCGCCACCTTTGTC

GTAGGAGGGCGCAGAGGTTGTGGT

GTATAGGGTCTGACGCCTGCCCGGTGCCGGAT

GTATGTCGCGTGAAGGACGTCCAAGAAAAGCC

GTGGATGTATAGAGTCTGACGCCTGCCCGTGG

TAAAGAGTAACGGAGGAGCCCAATTGGTACCC

TAACGAGCATACTGAATTATAGAGTGAATTCT

TATCAGCTTCCGACGGTAGGATATGGGCCTTG

TCCCGTGGATTGGGAGAAGTTTGAGTAAGAGC

TCGTTTCCCGGCTGATGCACCATGGAATTTCT

TCTGGCGACCATATCCGAAAGGAAATACCTTG

TGAGTATCAATTGGAGGGCAAGTCTGGTGTGG

TGCATTGCGCTCTTGGGATATGCC

TGTAAGCTTAGGAGCTGTATACAGAAAAAAAA

TGTCCGTTCGAGTCGGACCGGGGGCACCATGG

TGTCGTGAGATGTTGGGTTAAGTC

TGTGGGTTCAAGTCCCACCGTCGGCTCCATGG

5.3.3.6 DS (Dark stress)

AAAGACTCCTAGATCCTGATGCCTTTAATCCT

AAGATTCTGGGTTCGGGTCCCAGTGGGCGCTC

AAGGGGAATGTTGGACGTACCTTGACTGGCGG

AATGTGGCGTAGAAGTTAGCGTATTTGGTTTG

ACGGAAGGATTCTGATGGCACCTCCATGTCGG

ACTTTTGATTTGTGGTTCGCCGGGGATAACTG

AGCTCCTACGGAAGGATTCTGATGGCACCTCC

ATACATCGTTGCCATCGATGTGACCCGGGTTC

ATCTGATCTCCTAGATCGGATTGAGGGCCTGG

ATGGCCCCTGGTTTCCGGGTGGCCGGACCTGG

ATGGGCTTTCCCGGAAGGCACCGCCCGAGGCG

CACCCGAGCGACCCGGGTTCGTGTCCCGGTGG

CACCTCCATGTCGGCTCATCGAAAACCCATGG

CATTACTGTCAACTATGATACTTTCTTGACCC

CCAAAATAAGACGCGCCGGCCATCTGCCCTTG

CCGGAGATTCCAGGTTCGTGTCCTGGGCAGCT

CCGGTGGTCGTTTCCCGGCTGATGCACCATGG

CCTCAAGGTCGGGAGTTCGTGCCTCCCTGGGA

CCTGGAGGTAACAAAGAATCCTGGTAATCGAG

CCTGTCGCGCGGAAGACCCGGGTTCATTTCCC

CGAAGGCCATTTGACTCTAGAGCAATCAGGTC

CGAGAGAGATGTCGTAAGTTGTCTTTATGAGA

CGCCAGTTCGTTCCTGGCCAGGTGTACCATGG

CGCGGAAGACCCGGGTTCATTTCCCGGCGGCG

CGCTATCAGTGACGCTTAGCGGAGGTGCGCTG

CGTAAGTTCGCGAGAAGGGAGCCCCGGGCTTG

CTGACTGCAGATCAGCAGGTCCCTGGTTCAAA

CTGATGGCACCTCCATGTCGGCTCATCGAAAA

CTGGTTCAAATCCGGGTGTGCCCTCCA

CTGTTCTTATCAGTGTGACAACTG

GTAGGTGGCGTATTTGGTTTGGGTCCAAATGG

GTAGTGGTATCACATCCGCTTTGCGTGCGGGA

TATTAATTCTGGCAGCATTTTGGCATTGGGCA

TCACCATTGGTACTATTGATCAAAAGACCCTT

TGCTTGTGCCCCGGGTTCGGCTCCCGGCATGG

TGGGCGGCGGATGTCTTGCGGATGGTTCCTTC